Cray XT Systems

John Levesque

Dir. Cray Supercomputing Center of Excellence

Since We Last Met in San Francisco…

CRAY
North America
2008
TECHNICAL WORKSHOP

Upgrades & Shipments…

- We upgraded over *300* XT cabinets to XT4 quad core technology
- We released the first instance of the “Cray Linux Environment” and used it to achieve acceptance of “Franklin” and HECToR.
- We shipped our 1000th Cray XT cabinet on September 4th *(and 100% are still in service)*

Sandia National Laboratories
Red Storm System

- We completed our 3rd major upgrade to Red Storm
- Center section has been upgraded to Cray XT4 quad core modules
- Now running at 284 TFLOP peak performance

> ***"Red Storm is enabling us to carry out unprecedented simulations. For example, we are able to resolve climate calculations at 1/10 of a degree, which is well beyond the current state of the art. We are able to carry out an order-of-magnitude larger simulations, an order-of-magnitude faster than on any of our previous capability systems."***
>
> \- Bill Camp, Sandia Director of Computers, Computation, Information and Mathematics

Photo Courtesy of Sandia National Laboratories

Red Storm Supercomputer helps U.S. Navy Shoot Down Errant Satellite

The Science Challenge: Successfully Shoot Down a Failed Satellite

- Shoot down an errant satellite with a single missile strike.
- Determine the optimal hit point for destruction, minimizing the spread of debris
- Ensure that one missile strike would suffice.

HPC Challenge: Perform Complex Simulations to Optimize Strike

- Ran hundreds of impact simulations to answer critical technical questions affecting early decisions

Cray's Contribution

- The entire Red Storm system dedicated for about two months for simulations and planning the complex missile strike.
- Helped the Department of Defense (DoD) plan and execute the operation, as well as conduct follow-up analysis.

> "The architecture of the Red Storm XT system…was critical in facilitating the high-fidelity simulations required to provide confidence in a spectrum of scenarios to DoD," said James Peery, director of Sandia's Computer and Computation Sciences Center.

National Energy Research Computing Center (NERSC) Lawrence Berkeley National Laboratory

- "Franklin" installed and accepted in 2007
- Competitive procurement based on NERSC Sustained System Performance (SSP) metric
  - Over 19 TFLOPS measured *sustained* performance
- System upgraded to Quad Core, 350 Tflops in 2008
  - System sustained performance has doubled

"The Cray proposal [Cray XT4] was selected because its price/performance was substantially better than other proposals we received, as determined by NERSC's comprehensive evaluation criteria of more than 40 measures."

Bill Kramer
General Manager
NERSC Center

NERSC

US DoD Program

- Four of five systems for TI-08
- One of the largest DoD HPCMP system awards to a single vendor
- Four Cray XT5 systems to be located at top military research centers, including
  - Army Research Laboratory
  - Naval Oceanographic Office
  - Arctic Region Supercomputing Center

MRAP = Mine Resistant Ambush Protected

Alaska Regional Supercomputing Center

- 35 Tflop, 5 Cabinet XT5 system
- System is named "Pingo" after the permafrost formations that can grow to over 200 feet high
- TDS system is named "Ognip", which is a collapsed Pingo.

Finnish IT Center for Science (CSC-Finland)

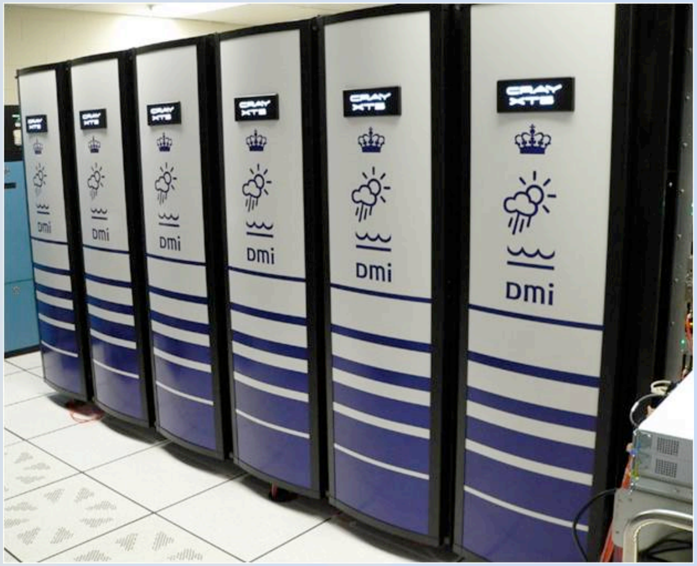
- We installed our first "XT9" (XT4 +XT5)
- Research Areas
  - Physics
  - Chemistry
  - Nanotechnology
  - Bioscience
  - Applied Mathematics
  - Engineering

"We selected the Cray supercomputer after an extensive acquisition process that involved surveying 35 different research groups, closely analyzing the available technologies and benchmarking competing systems."

Kimmo Koski
Managing Director
CSC-Finland

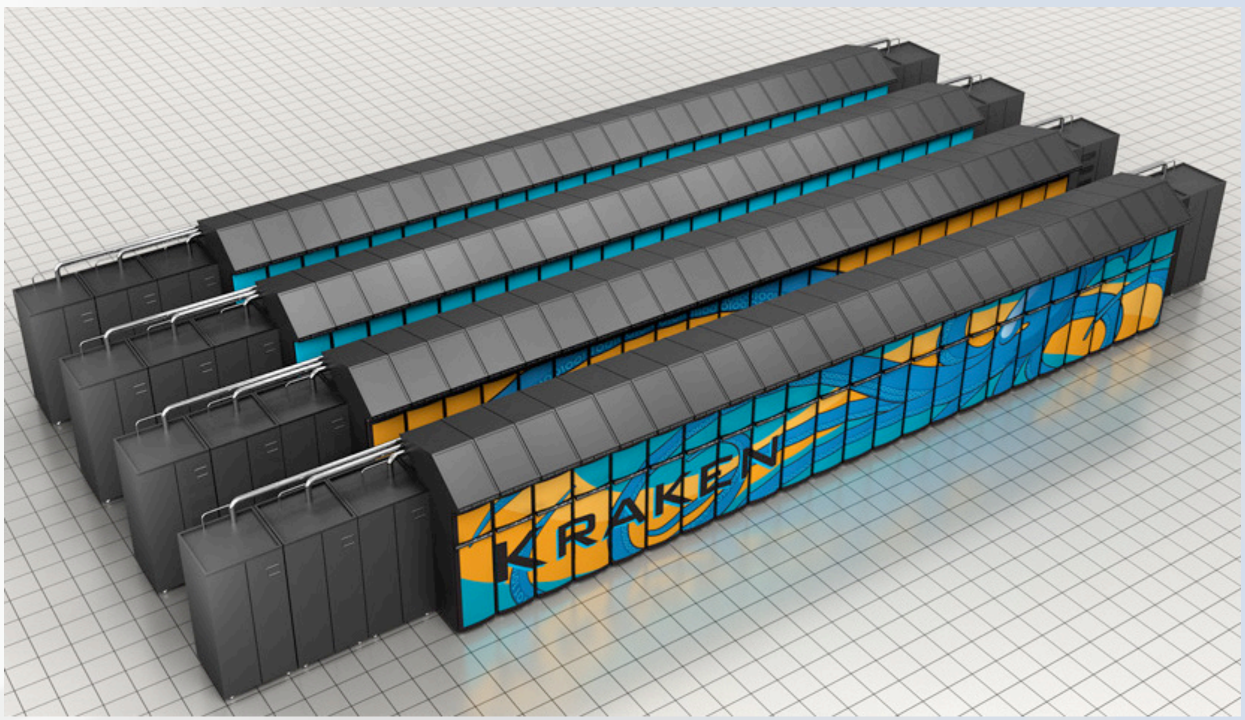
We Installed a 2nd Production Weather Site

- Two 3-cabinet systems
  - One for operational NWP
  - One for research
  - Full failover
  - External Lustre
- Will significantly enhance DMI's NWP and climate assessment capabilities
- This is also our first external Lustre implementation in Europe



National Science Foundation – Track 2

- Awarded to the University of Tennessee
- Cray XT4 followed by near petaflop Cray XT5
- Housed at the University of Tennessee – Oak Ridge National Laboratory Joint Institute for Computational Sciences




KRAKEN

ORNL Petaflop System

Jaguar: World's most powerful computer. Designed for science from the ground up

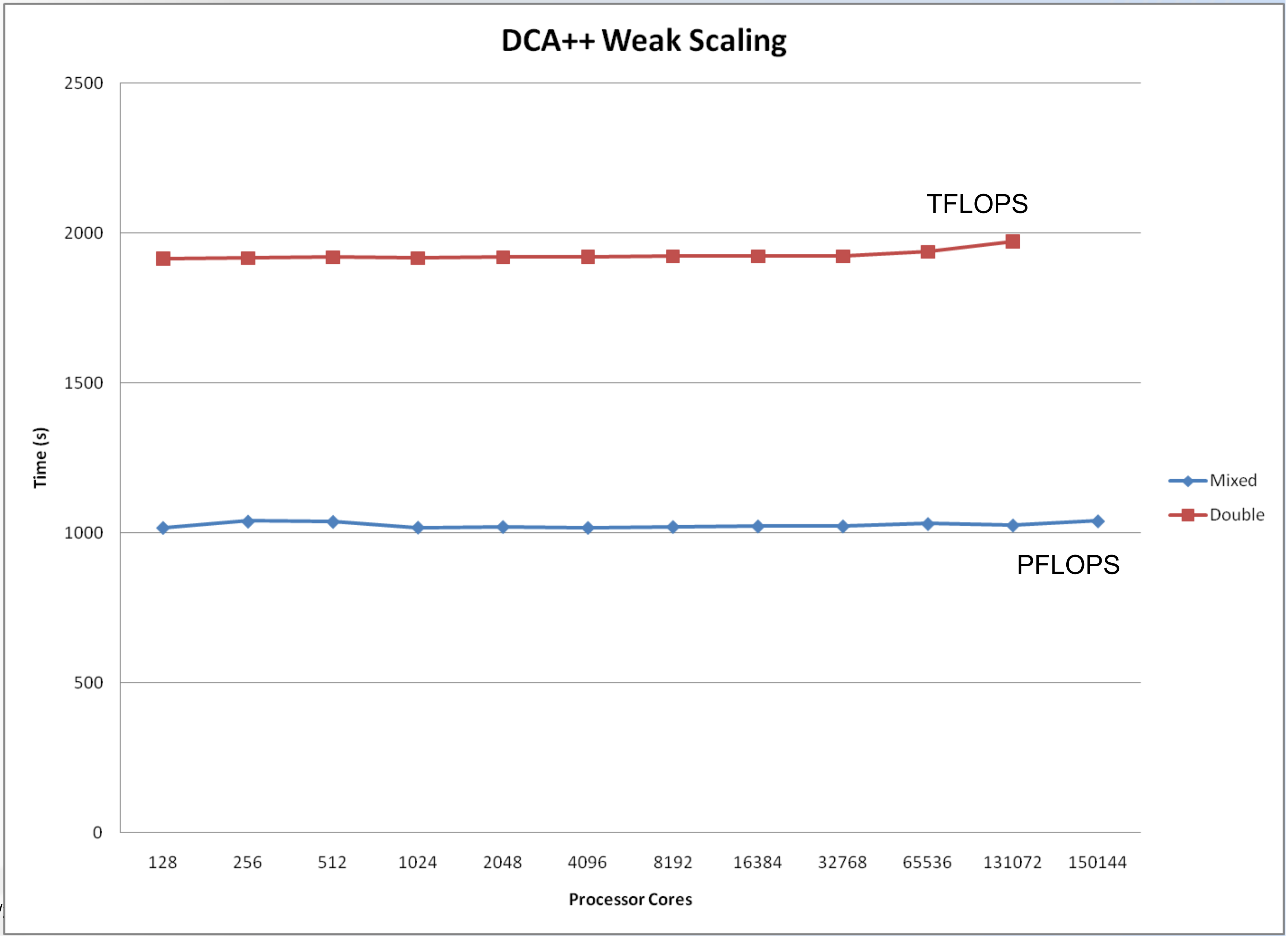
Peak Performance	**1.38 Petaflops**
System Memory	300 Terabytes
Disk Space	10.7 Petabytes
Disk Bandwidth	240+ Gigabytes/second
Processor Cores	150,000

Short History of Jaguar Petaflop

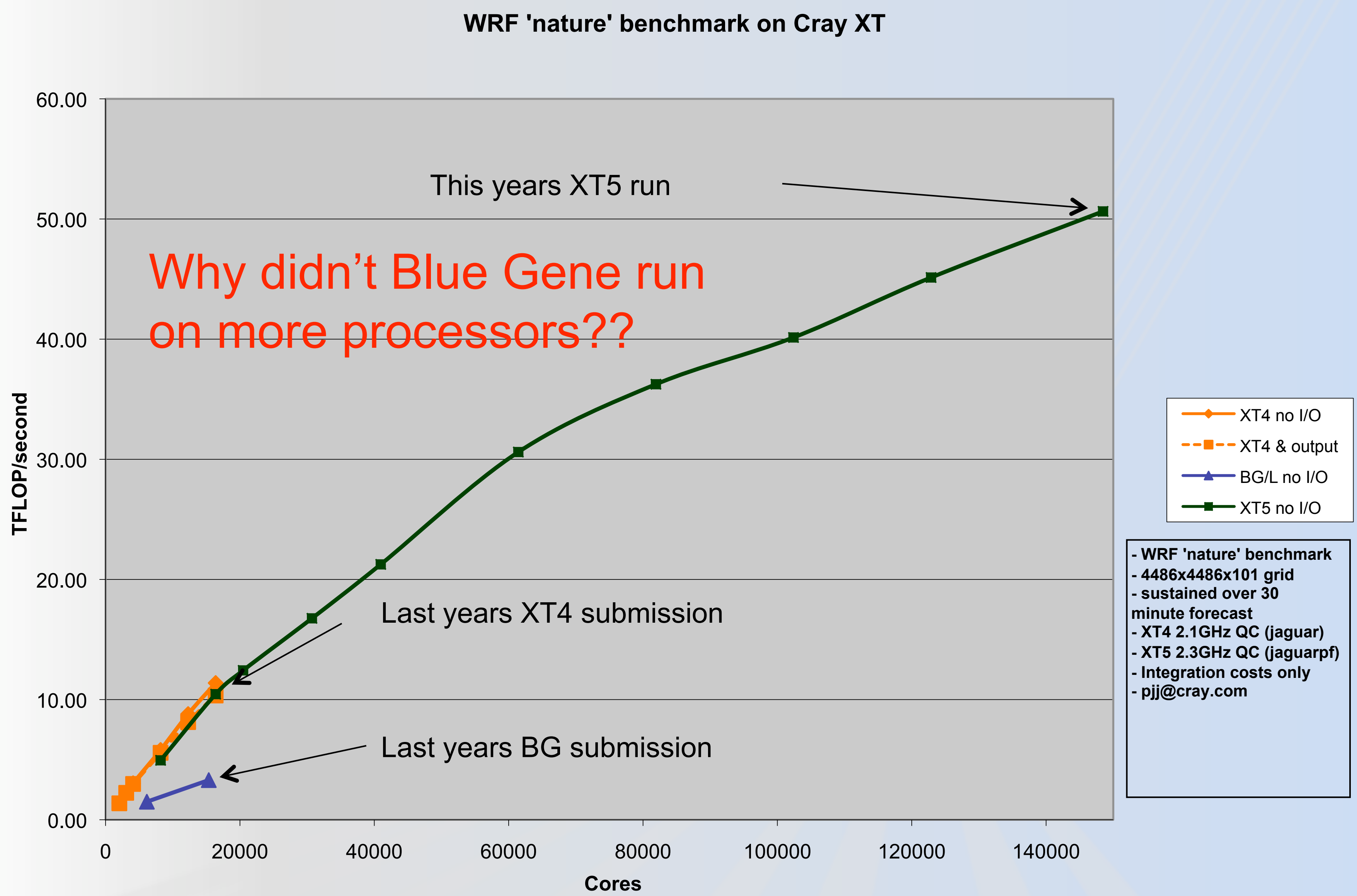
- System completely installed on the floor at ORNL the first week of October. 200 new liquid cooled cabinets
  - Amazing ramp up of Cray manufacturing to make this happen
- Used HPL to shake down the machine
  - Adrian Tate is now a Tennessee folk hero and he is British
  - Get a long enough run on a brand new system to place high in Top 500
- Ran a few applications between HPL runs
  - Amazed at scalability
- November 7th we started running applications
  - Eight applications, from 5 different science areas set World records on performance
- November 10th SPECfem3d ported and ran on 149784 cores
  - Actually hit a code error, because a problem of this size and resolution had never been tried before. This was fixed quickly and the run succeeded.
- November 12th Lin-Wang Wang code from UCB ported and run on 149144 cores
- Steve Whalen made numerous HPCC runs this last week

Early Science Applications

Science Area	Code	Contact	Cores	% of Peak	Total Perf	Notes	Scaling
Materials	DCA++	Schulthess	150,144	97%	1.3 PF*	Gordon Bell Winner	Weak
Materials	LSMS/WL	ORNL	149,580	76.40%	1.05 PF	64 bit	Weak
Seismology	SPECFEM3D	UCSD	149,784	12.60%	165 TF	Gordon Bell Finalist	Weak
Weather	WRF	Michalakes	150,000	5.60%	50 TF	Size of Data	Strong
Climate	POP	Jones	18,000		20 sim yrs/ CPU day	Size of Data	Strong
Combustion	S3D	Chen	144,000	6.00%	83 TF		Weak
Fusion	GTC	UC Irvine	102,000		20 billion Particles / sec	Code Limit	Weak
Materials	LS3DF	Lin-Wang Wang	147,456	32%	442 TF	Gordon Bell Winner	Weak

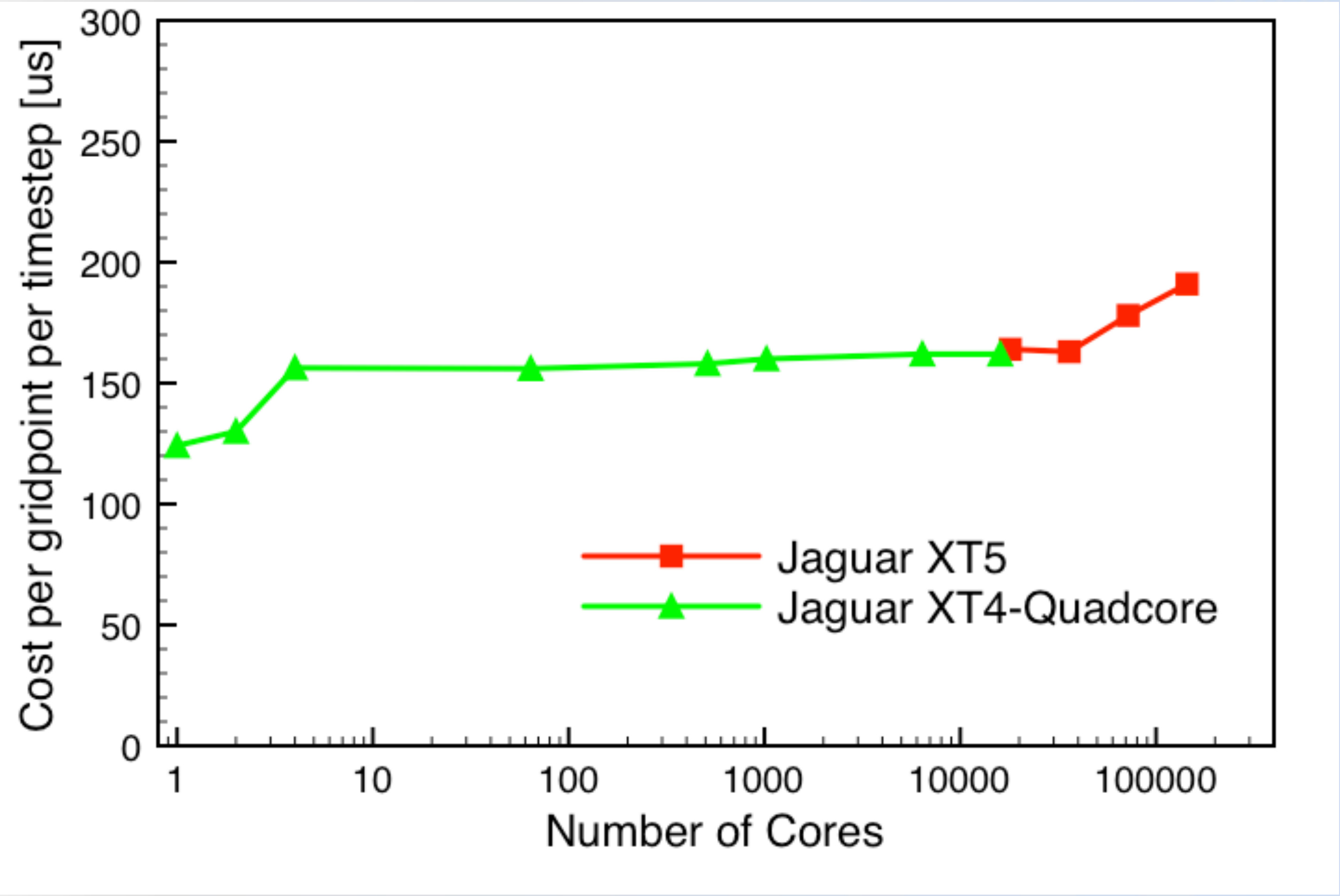


DCA++ Weak Scaling
2500
2000
1500
1000
500
0
Time (s)
TFLOPS
PFLOPS
Mixed
Double
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
150144
Processor Cores



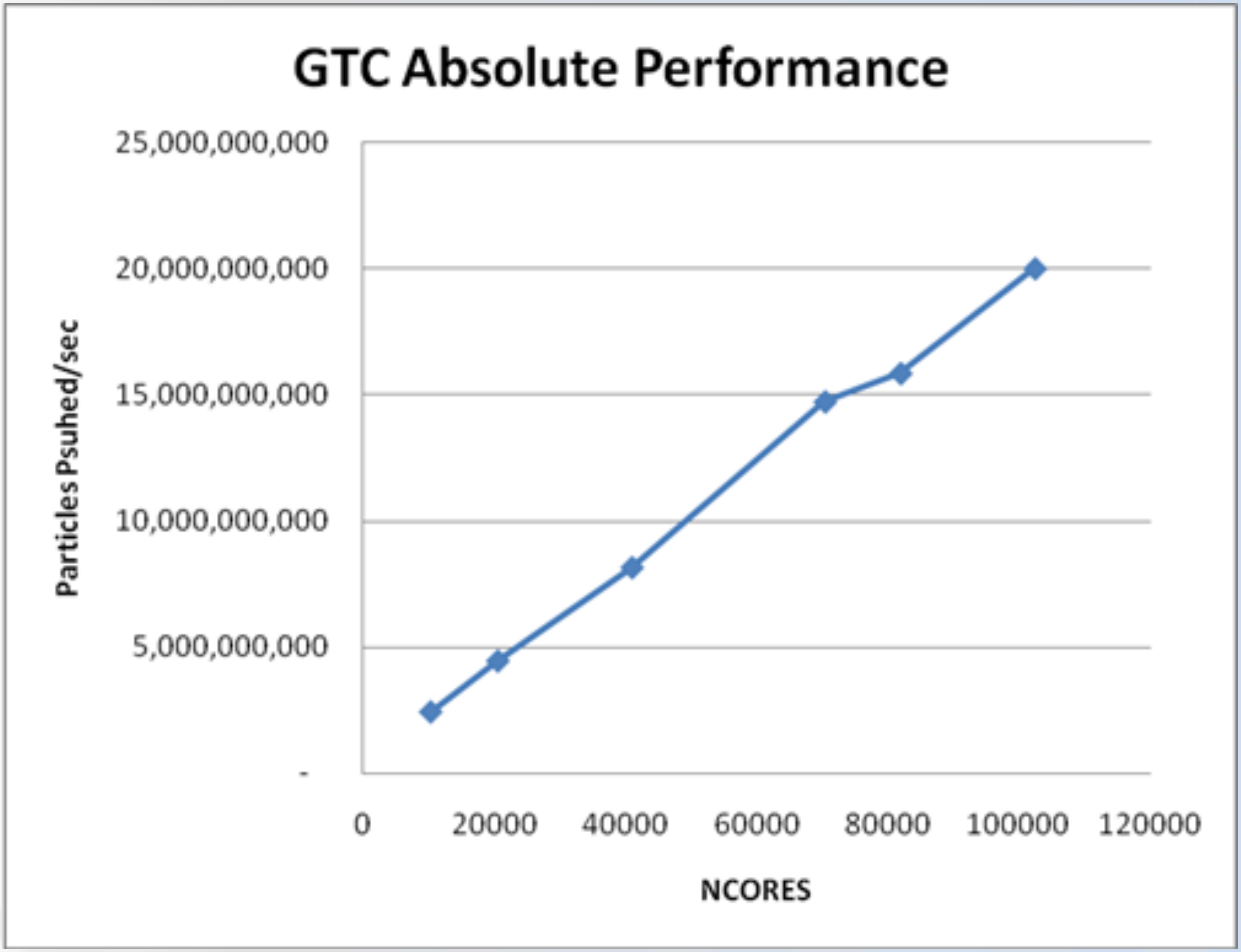
WRF 'nature' benchmark on Cray XT
This years XT5 run
Why didn't Blue Gene run on more processors??
Last years XT4 submission
Last years BG submission
TFLOP/second
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0
20000
40000
60000
80000
100000
120000
140000
Cores
XT4 no I/O
XT4 & output
BG/L no I/O
XT5 no I/O
- WRF 'nature' benchmark
- 4486x4486x101 grid
- sustained over 30 minute forecast
- XT4 2.1GHz QC (jaguar)
- XT5 2.3GHz QC (jaguarpf)
- Integration costs only
- pjj@cray.com

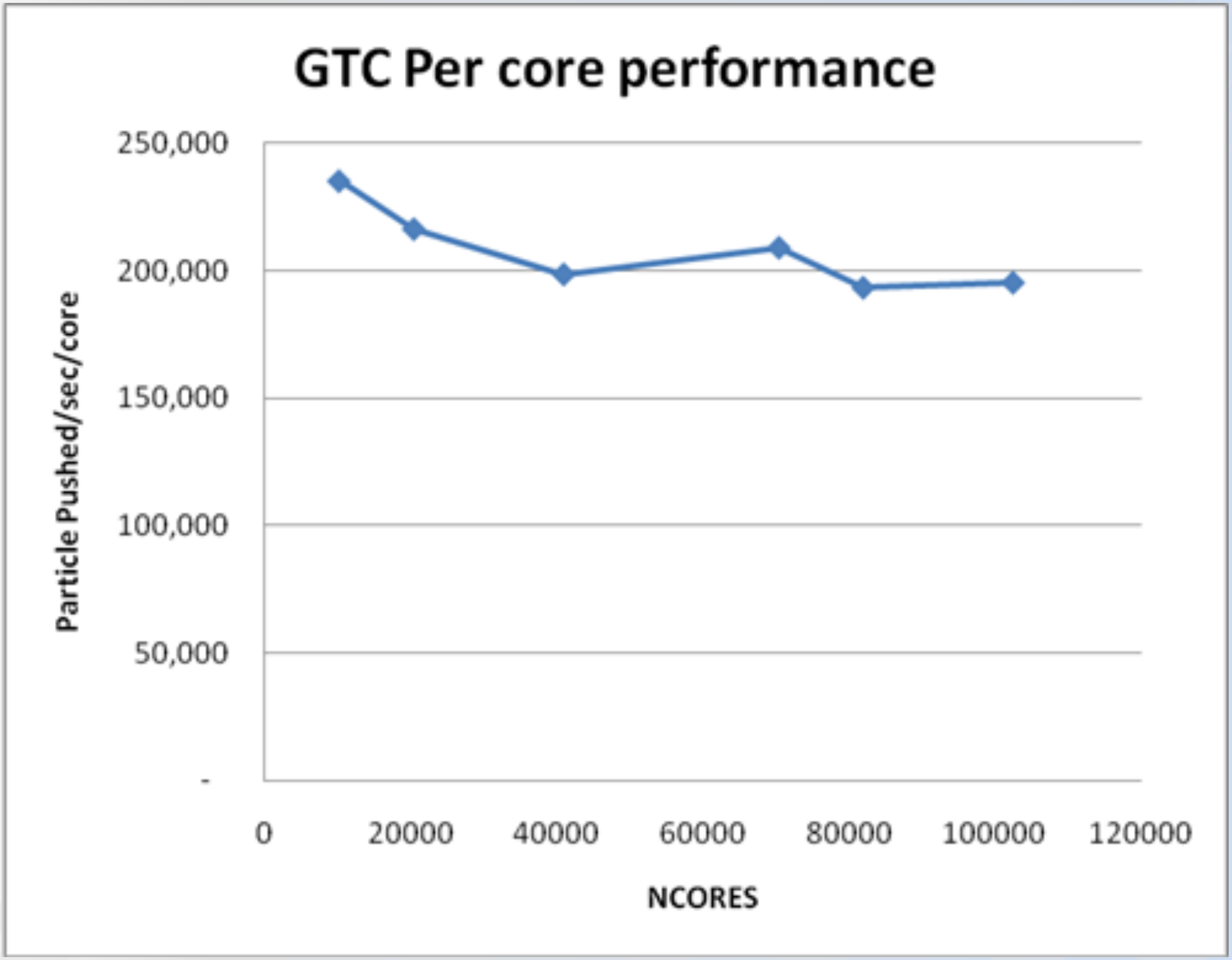
S3D-C2H4



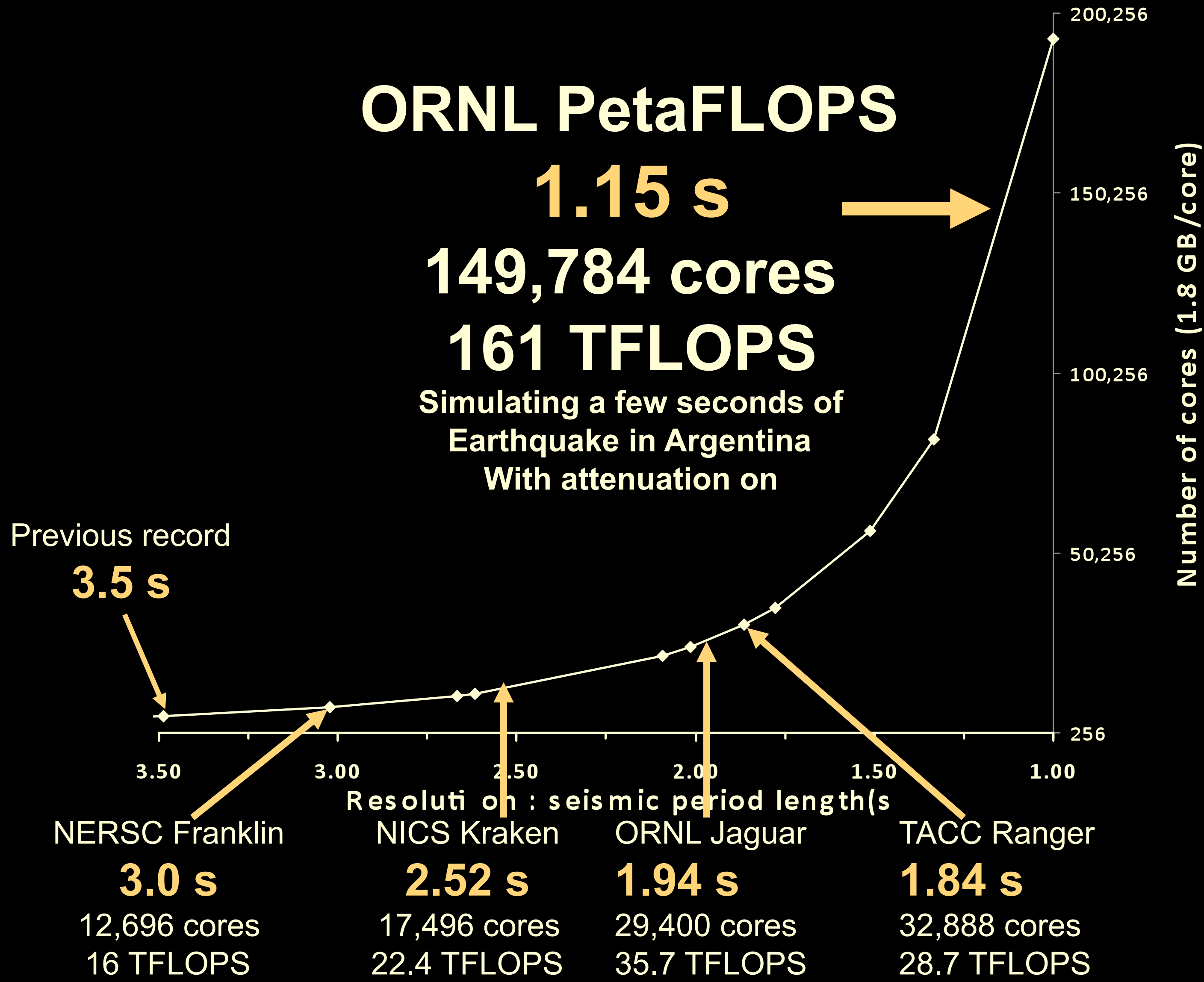
Cost per gridpoint per timestep [us]
300
250
200
150
100
50
0
1
10
100
1000
10000
100000
Number of Cores
Jaguar XT5
Jaguar XT4-Quadcore

Run at 142000 failed due to Code Limitations



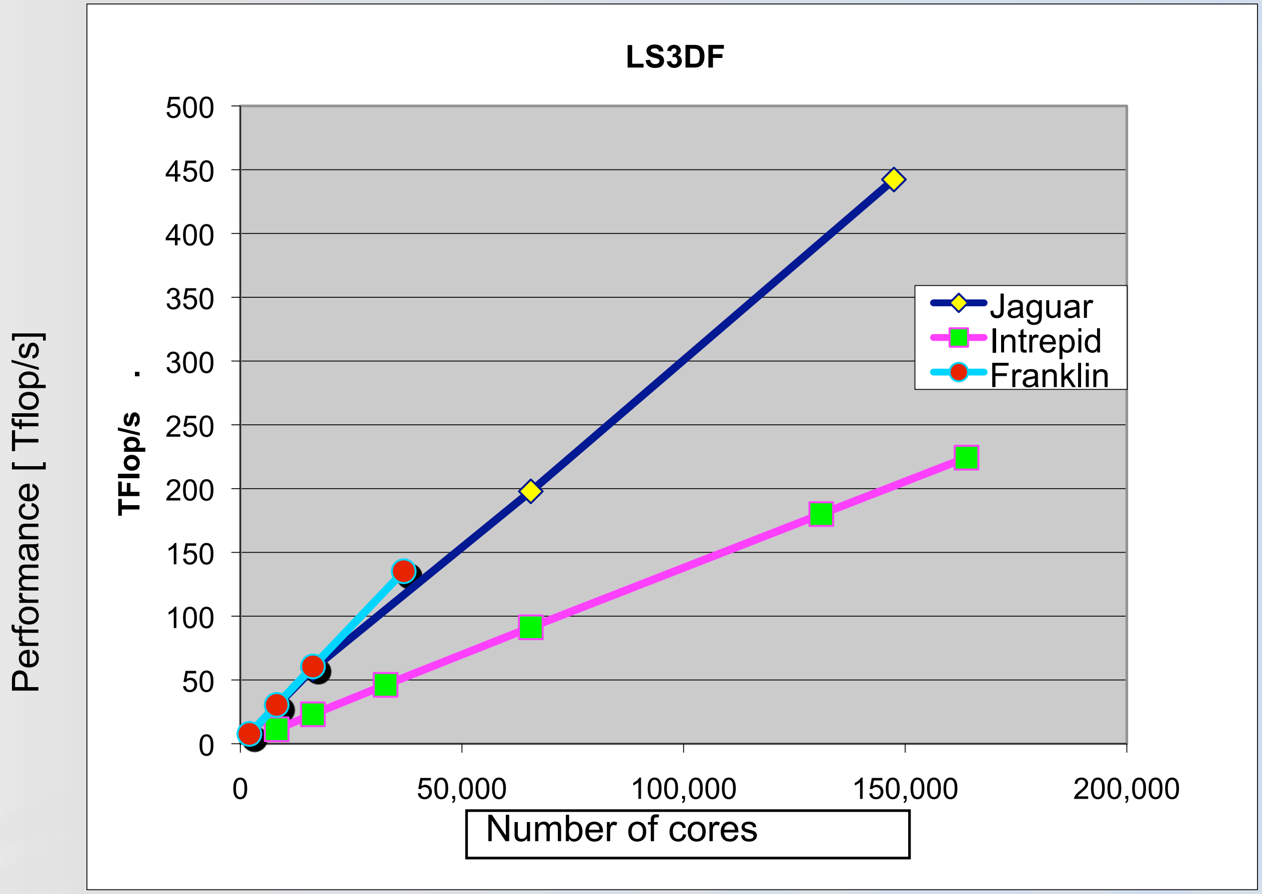

GTC Per core performance
250,000
200,000
150,000
100,000
50,000
-
0
20000
40000
60000
80000
100000
120000
Particle Pushed/sec/core
NCORES

Results - on the road to a PetaFLOPS system



ZnTeO alloy weak scaling calculations

- First large scale run on Franklin at NERSC: 135 Tflops, 40% efficiency
- Subsequent runs on Intrepid at ALCF: 224 Tflops, 40% efficiency
- Final runs on Jaguar XT5 at NCCS: 442 Tflops, 33% efficiency



Note: Ecut = 60Ryd with *d* states, up to 36864 atoms

Why are DoE Office of Science Researchers so well prepared for Petaflops at ORNL?

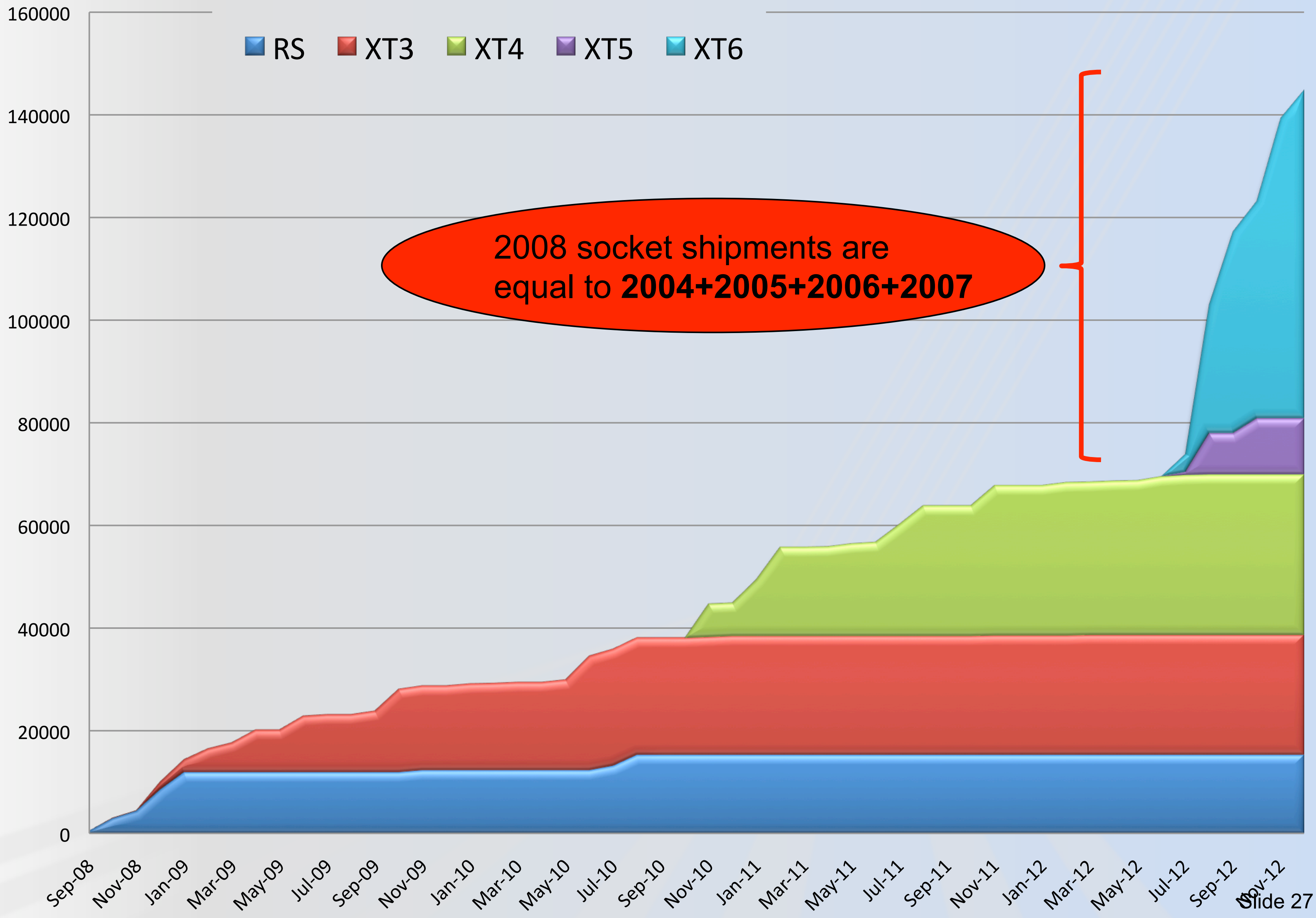
- For the past fours years Cray has worked with DoE Office of Science developers and the Computer Science Group at ORNL to prepare their applications for larger and larger processor counts
- They also have some of the best application developers in the industry

Shanghai Shipments

- We just shipped our first Shanghai-based system to JAIST
- JAIST was our first Cray XT3 customer outside of the United States
- The new system is 10 times more powerful than the previous
- System configured with ECOphlex liquid cooling

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Cumulative Socket Shipments



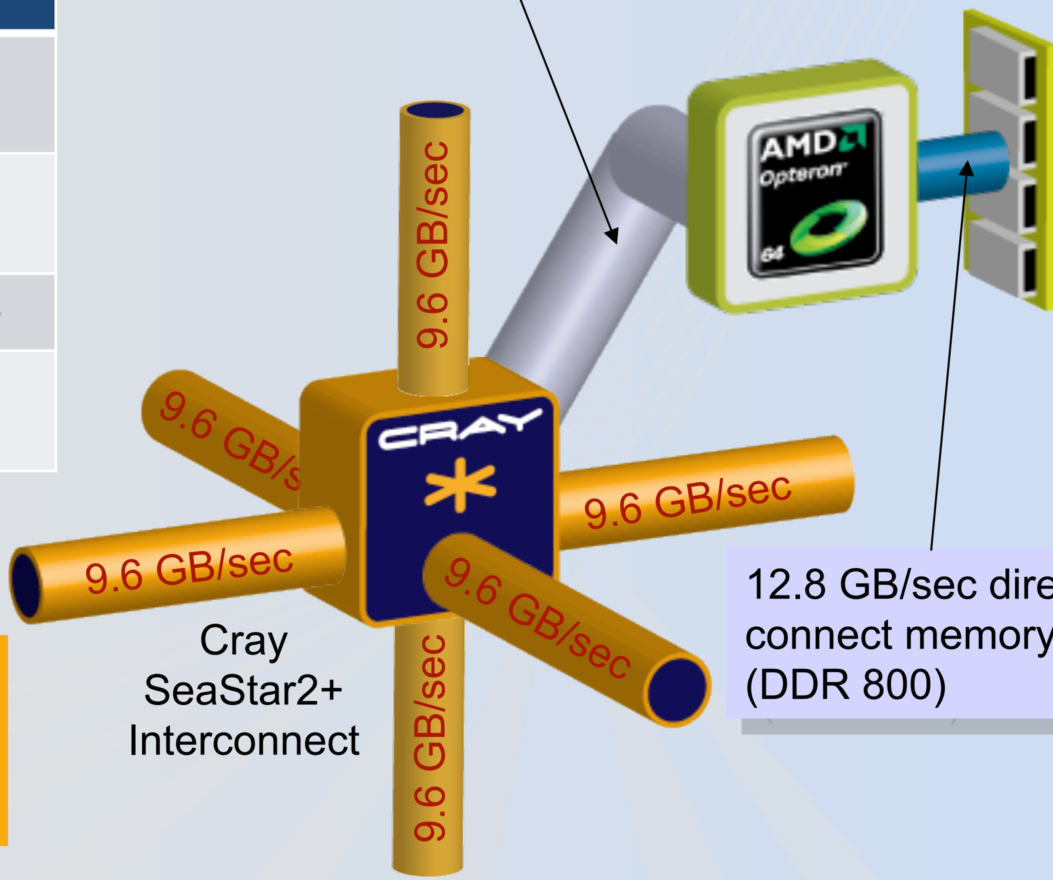
Cray XT5

Recall the Cray XT4 Node…

Cray XT4 Node Characteristics	
Number of Cores	4
Peak Performance	> 35 Gflops/s
Memory Size	2-8GB per node
Memory Bandwidth	12.8 GB/sec

2.3 GHz Budapest is the last processor upgrade for this product

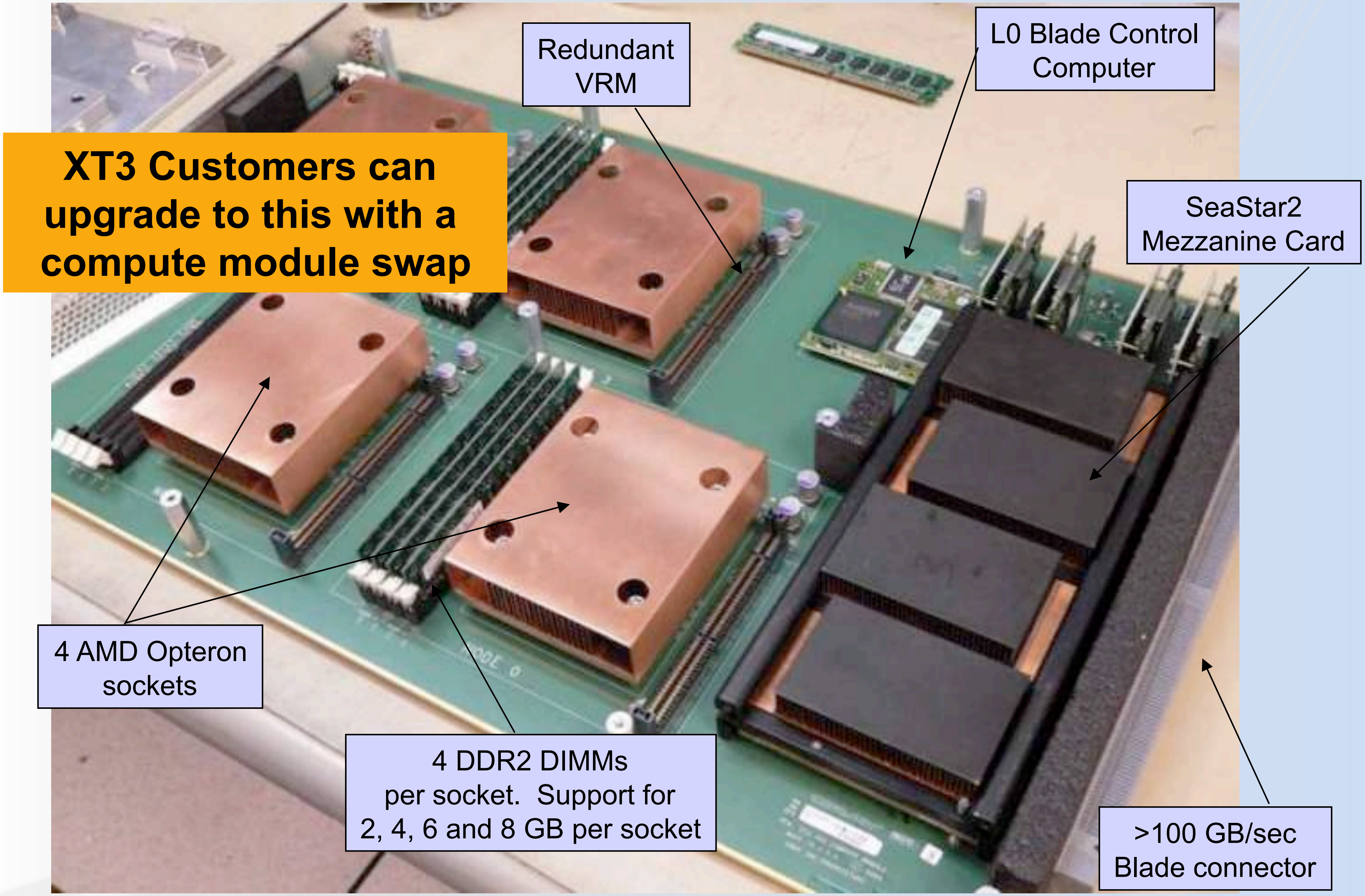
6.4 GB/sec direct connect HyperTransport



Cray SeaStar2+ Interconnect

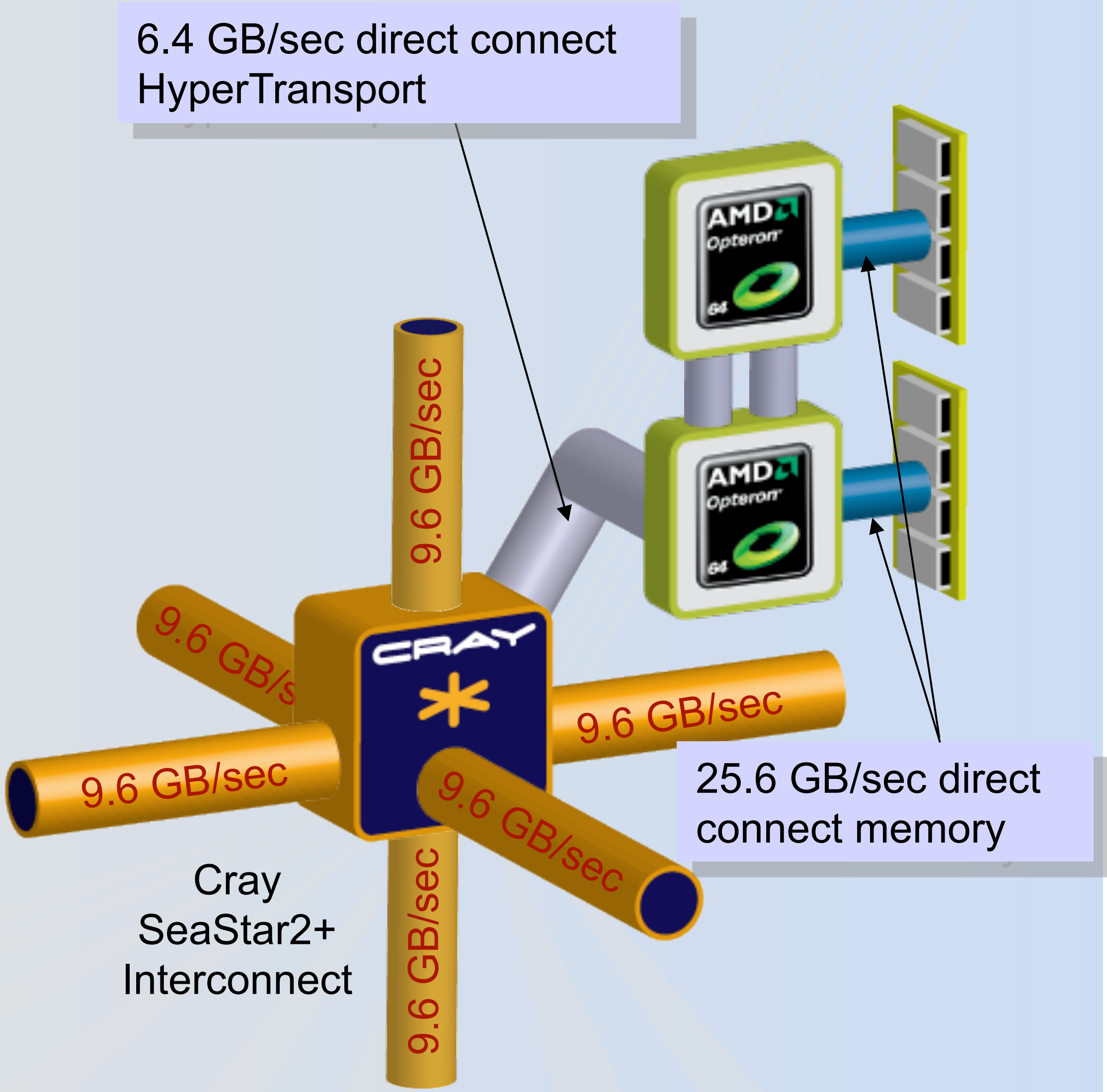
12.8 GB/sec direct connect memory (DDR 800)

And Packaging…

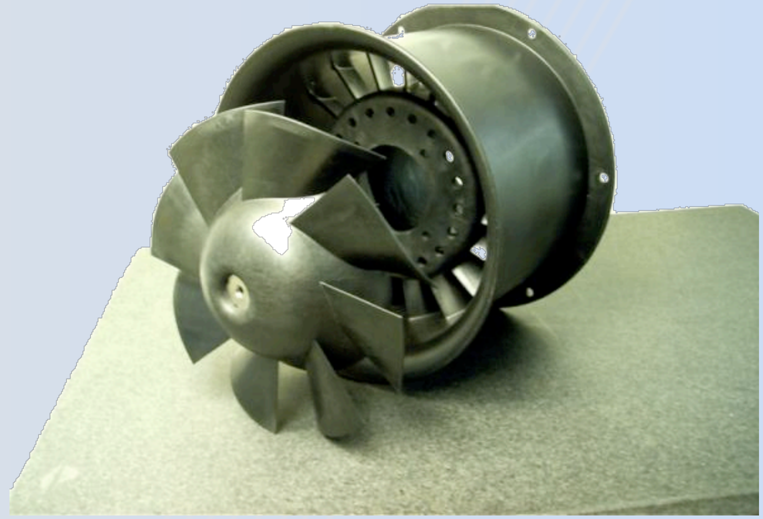


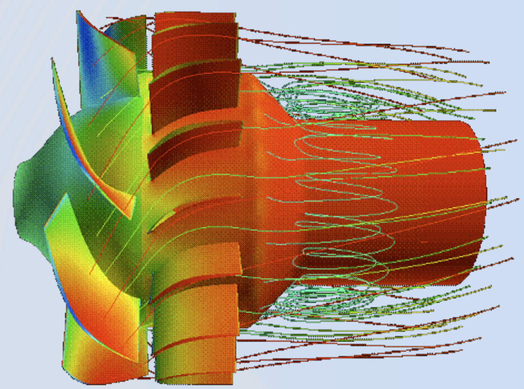
Cray XT5 Node

Cray XT5 Node Characteristics	
Number of Cores	8 or 12
Peak Performance Shanghai	76-86 Gflops/sec
Peak Performance Istanbul	125 Gflops/sec
Memory Size	8-32 GB per node
Memory Bandwidth	25.6 GB/sec

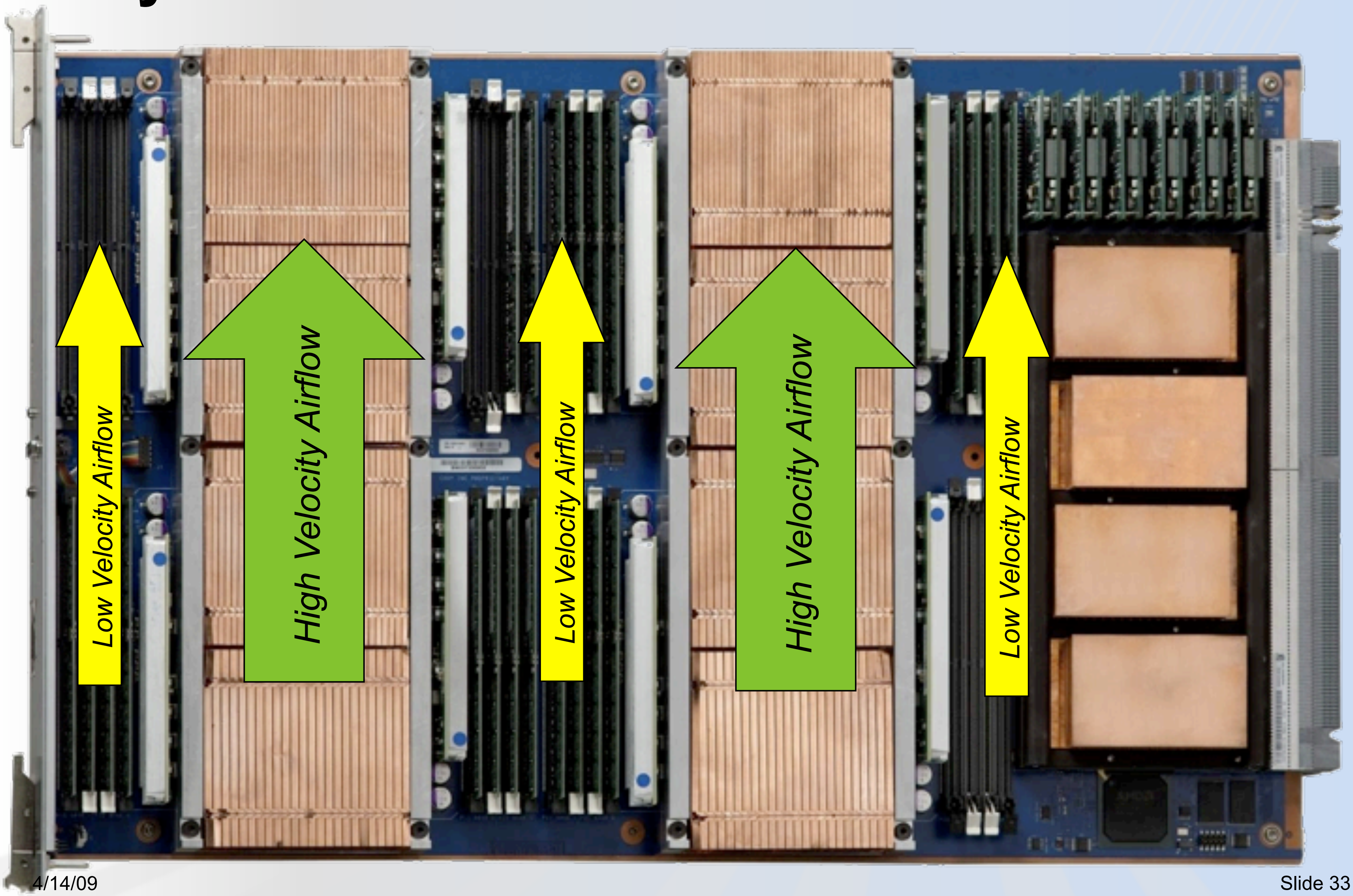


New Axial Turbofan – 78% Efficient

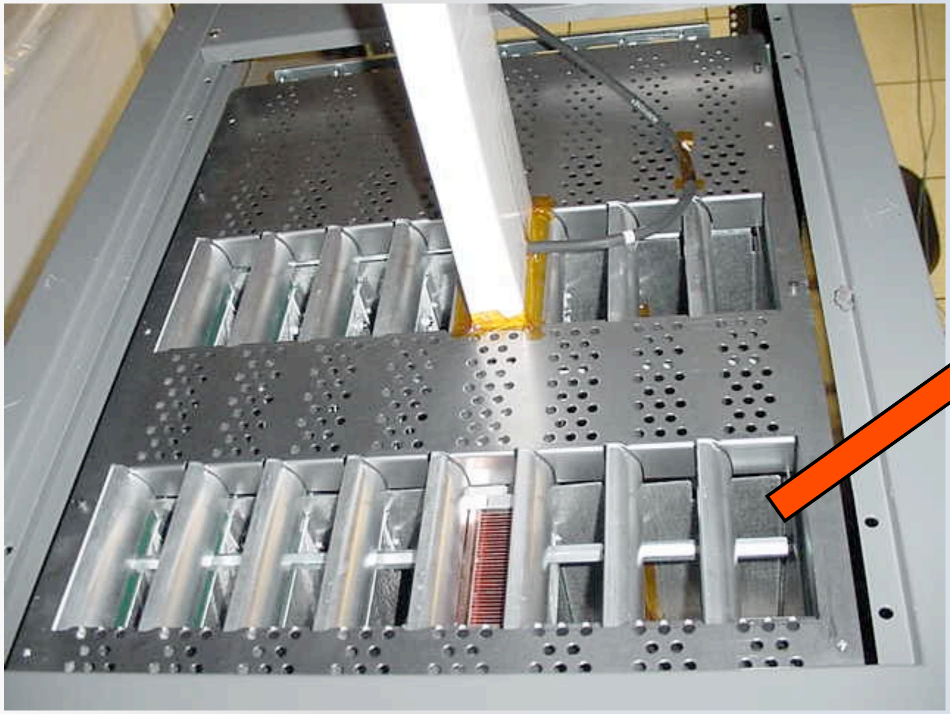


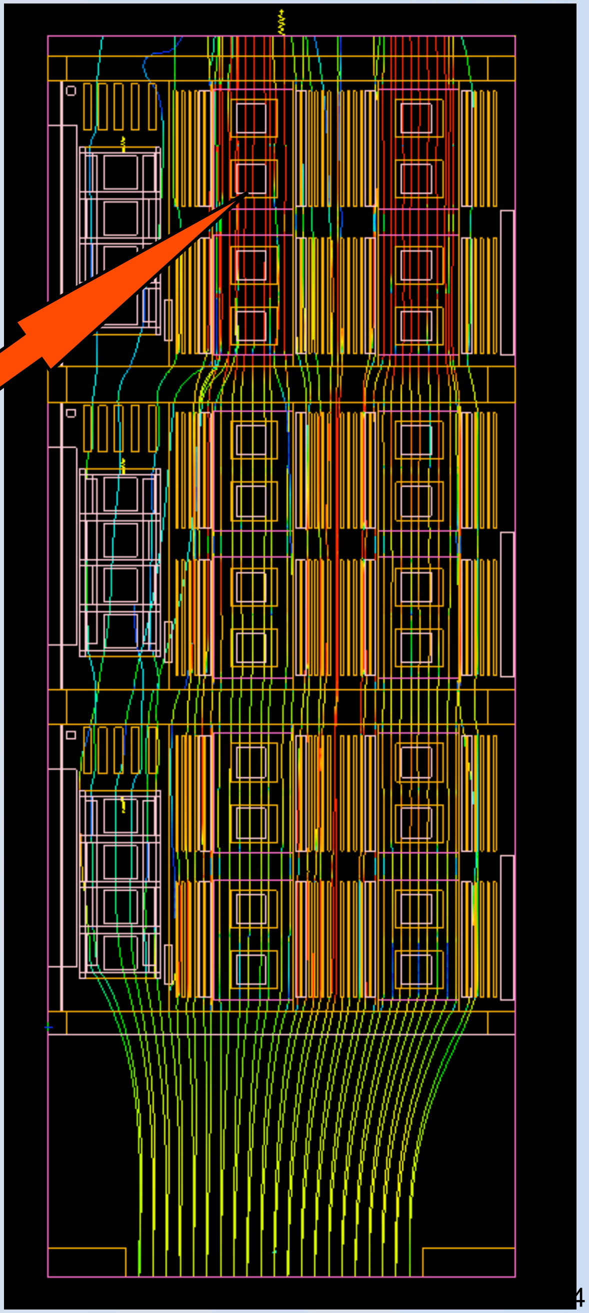


Cray XT5 Module



Chassis Air Flow Management

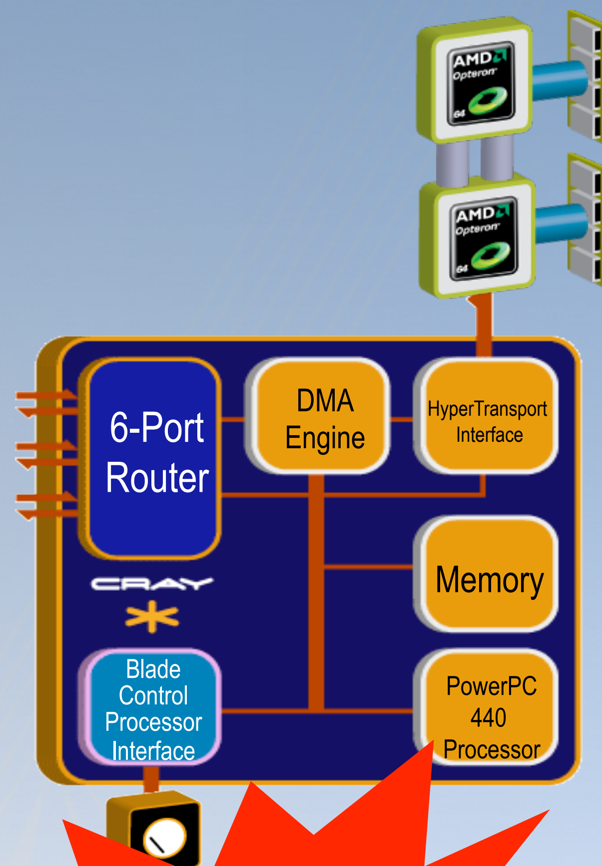




- Creates higher volume flow over processors
- Crossover from memory lanes between chassis delivers higher airflow to upper processors
- Graduated heat sink design maintains a 2 degree difference between the first and last processor in the air stream

Cray SeaStar2+ Interconnect

- New firmware was released with “Amazon” in 2008 that will improved SeaStar performance
- Improvements:
  - Improved packet arbitration and aging algorithm lowers global latency
  - Using 4 virtual channels improves sustained global bandwidth

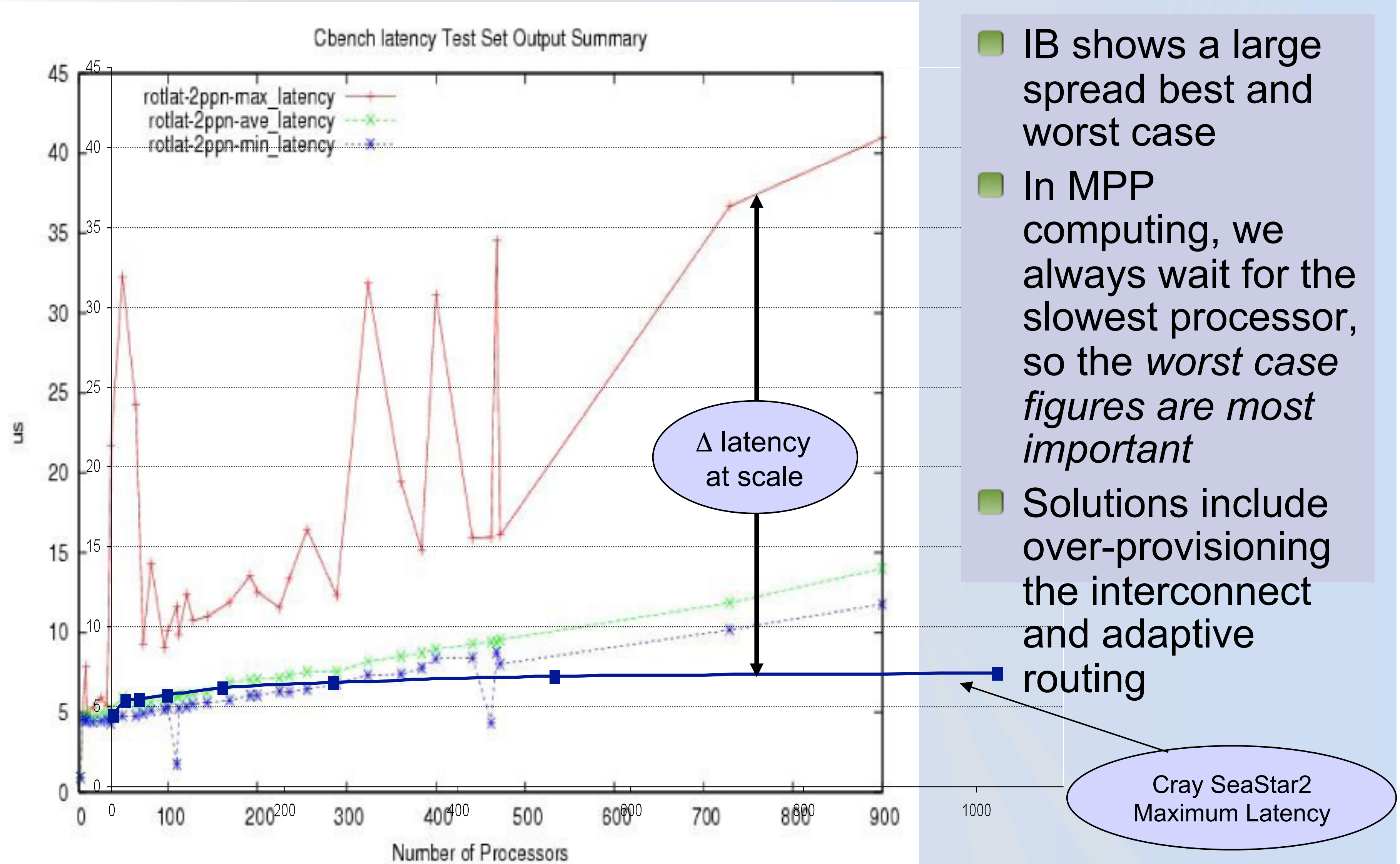


Packet arbitration and aging Improvement	
PTRANS	4.60%
MPIFFT	12.4%
AllReduce	12.4%
AllToAll	36.3%

Multiple virtual channels Improvement	
PTRANS	10–25%
MPIFFT	25%
RandomRing bandwidth	>40%

Now Scaled to 150,000 cores

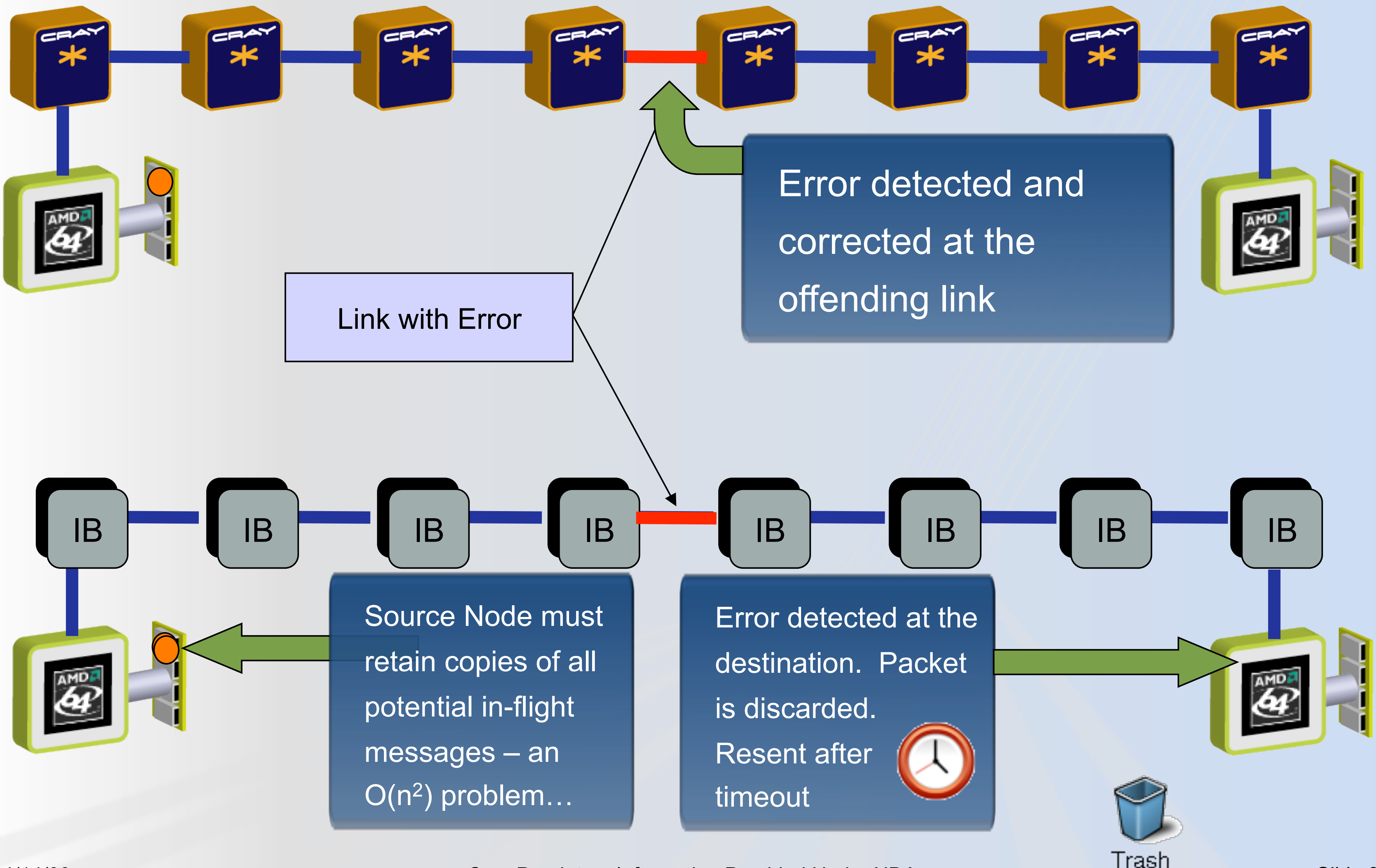
IB Cbench Latency



- IB shows a large spread best and worst case
- In MPP computing, we always wait for the slowest processor, so the *worst case figures are most important*
- Solutions include over-provisioning the interconnect and adaptive routing

Source: Presentation by Matt Leininger & Mark Seager, OpenFabrics Developers Workshop, Sonoma, CA, April 30th, 2007

The Importance of Link Level Reliability



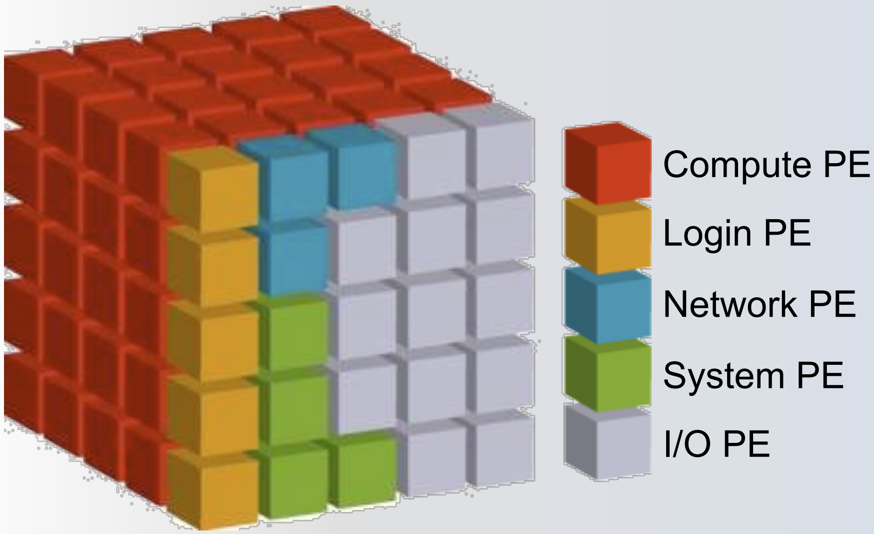
Cray XT5 Futures

- In mid 2009, we'll see a 6-core part from AMD code named "Istanbul" (Dual socket nodes up to 125 Gflops)
- In 2010 we'll be able to upgrade the interconnect to Gemini (more on that later)
- And new cabinets are upgradeable to the next compute blade…

Software

Scalable Software Architecture: Cray Linux Enviornment

"Primum non nocere"



Service Partition

Specialized Linux nodes

- Microkernel on Compute PEs, full featured Linux on Service PEs.
- Service PEs specialize by function
- Software Architecture eliminates OS "Jitter"
- Software Architecture enables reproducible run times
- Large machines boot in under 30 minutes, including filesystem

Software Perspective

- In the past year we:
  - Released CNL, and it's running on over 70% of installed cabinets
  - It's been scaled to 150,000+ cores
  - Global Arrays are working
  - UPC & CAF have been introduced on the XT

  - Improved MPT performance (shared memory device)
  - Updated performance tools,
    - Automatic profiling analysis
    - Performance measurement for OpenMP
    - Profile guided rank placement
  - Support for FFTW, PETSc & libgoto in Cray's scientific libraries
  - And we've released support for the X2 and XMT & XT5 !

  - More than doubled the software MTTI

Should Scientists be Computer Hackers

John,

There is not much really to compare to Ranger, it really is a difficult system. We struggle to get jobs to run. We have been debugging the system for over 6 months and finding work arounds to the buggy system. The only direct comparison we have is the single 32K job that ran which is : 28.7 TFlops(Ranger) vs 36.9TFLOPS (Cray)

That 32K job on Ranger only ran by luck, we submitted 10s of jobs all failing with memory leak problems from the system (not our code).

I know the TFLOPS number doesn't do your system justice in a comparison. The real comparison is it took a team of people months to try to get a single 32K job to complete on Ranger, while it took you only days to run multiple 32K+ and 150K
jobs on a pre-production system.

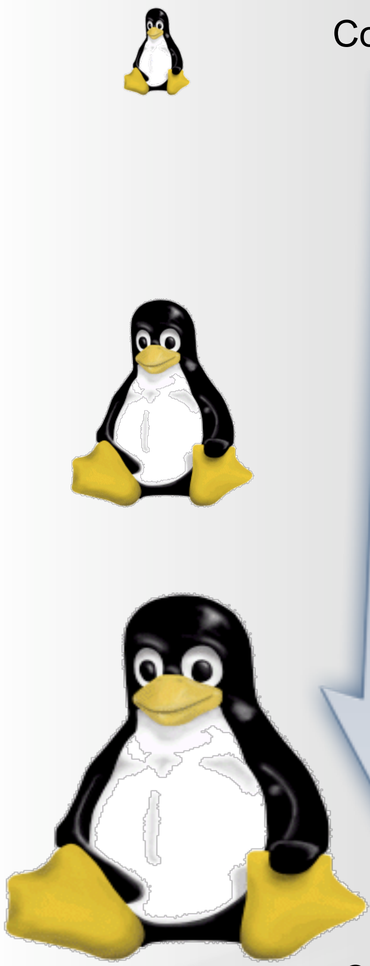
Laura

Laura Carrington of UCSD, Gordon Bell Finalist working on SPECFEM3D
Ranger is the University of Texas NSF system.

Cray Linux Environment

- Compute node kernels
  - XT CNL
  - XT Catamount
  - X2 CNL
  - XMT
- Service node kernel
  - Supports all compute nodes
- I/O
  - Lustre file system
  - DVS (Data Virtualization Service)
- Networking
  - Portals
  - TCP/IP
  - Gemini Network interface and Distributed Memory API
- Operating system services
  - Checkpoint / restart
  - NodeKARE (Node Knowledge And Reconfiguration) aka Node health Checker
  - CSA (Comprehensive System Accounting)
- System management
  - Interface to system data
  - ALPS (Application-Level Placement Scheduler)
    - Interfaces to PBS Pro, Moab/Torque and LSF
  - Command interface

Cray Linux Environment “The Vision”



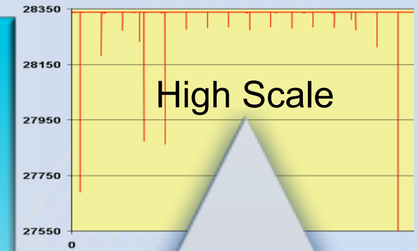
Less Compatibility

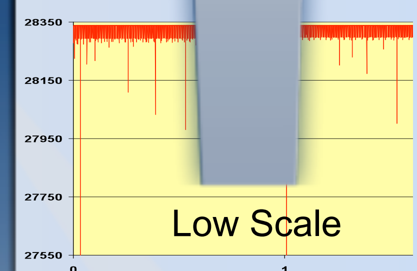
Full Compatibility

Capability (Ultra-light Linux Image)

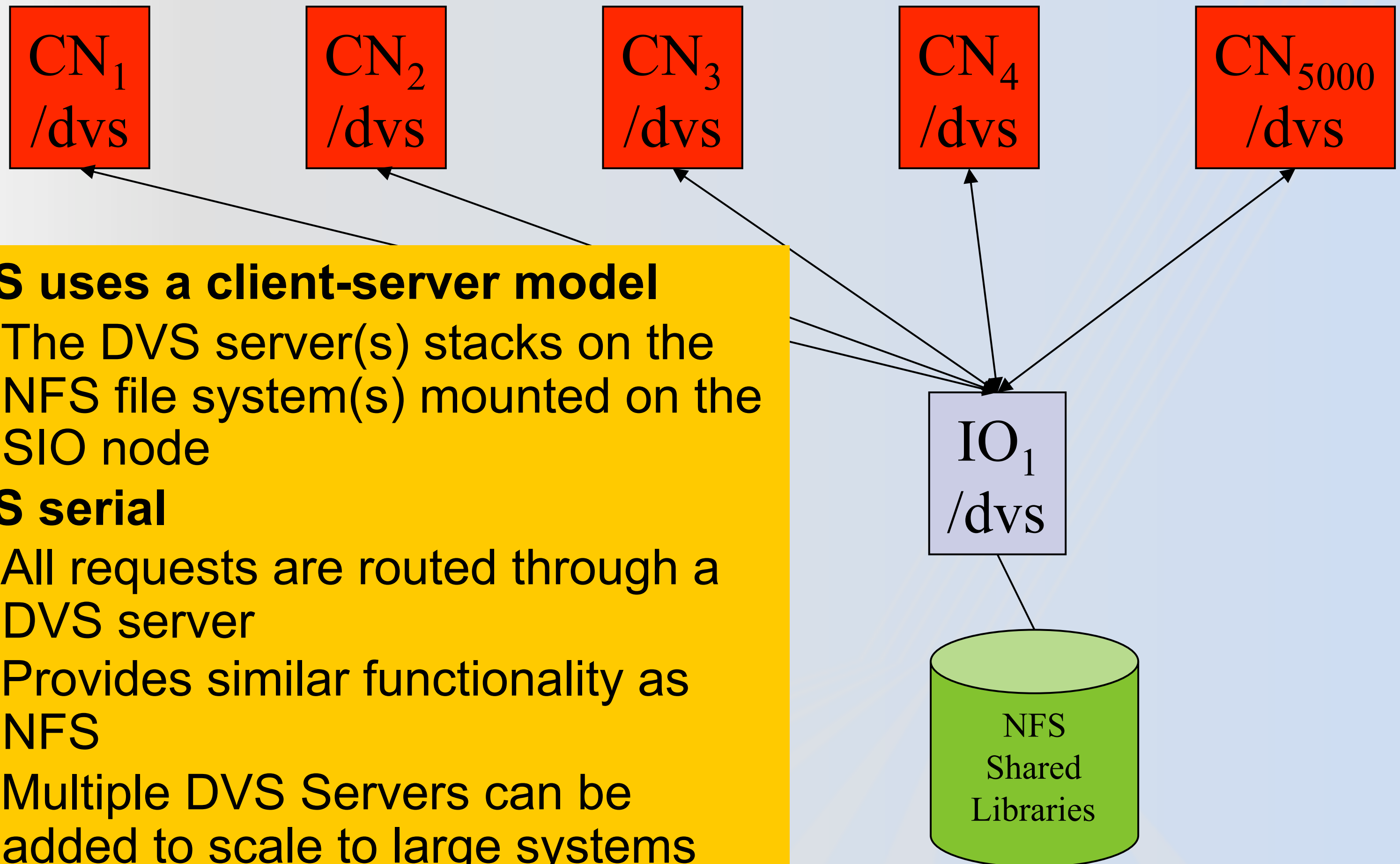
Capacity/Production
(Mid-weight Linux Image)

Shrink-wrap
3rd Party Application
(Full Linux Image and all services)





Scaling Shared Libraries with DVS



- **DVS uses a client-server model**
  - The DVS server(s) stacks on the NFS file system(s) mounted on the SIO node
- **DVS serial**
  - All requests are routed through a DVS server
  - Provides similar functionality as NFS
  - Multiple DVS Servers can be added to scale to large systems
  - DVS Server can be a “repurposed compute node

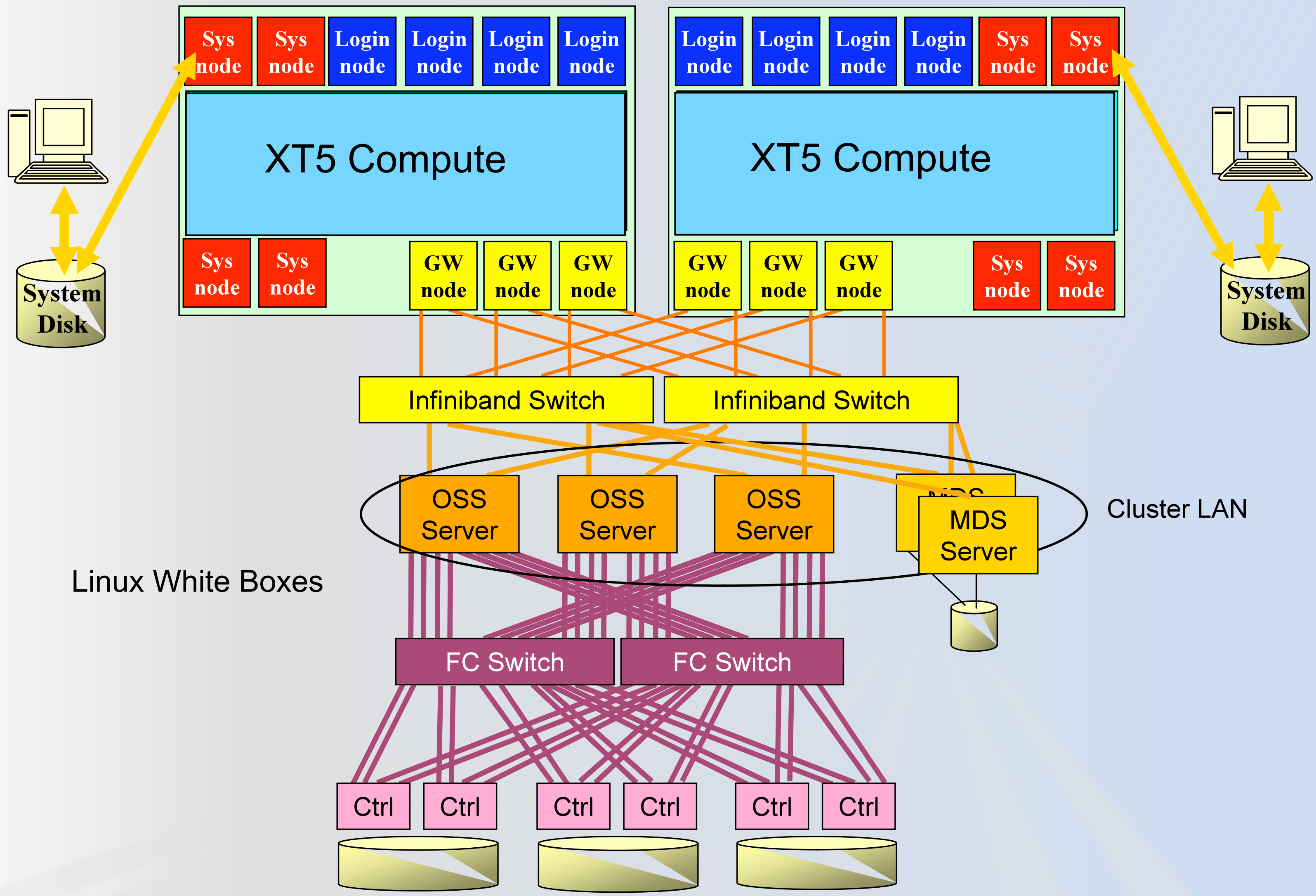
Programming Environment

will be covered by Luiz DeRose, Adrian Tate and Nathan Wichmann

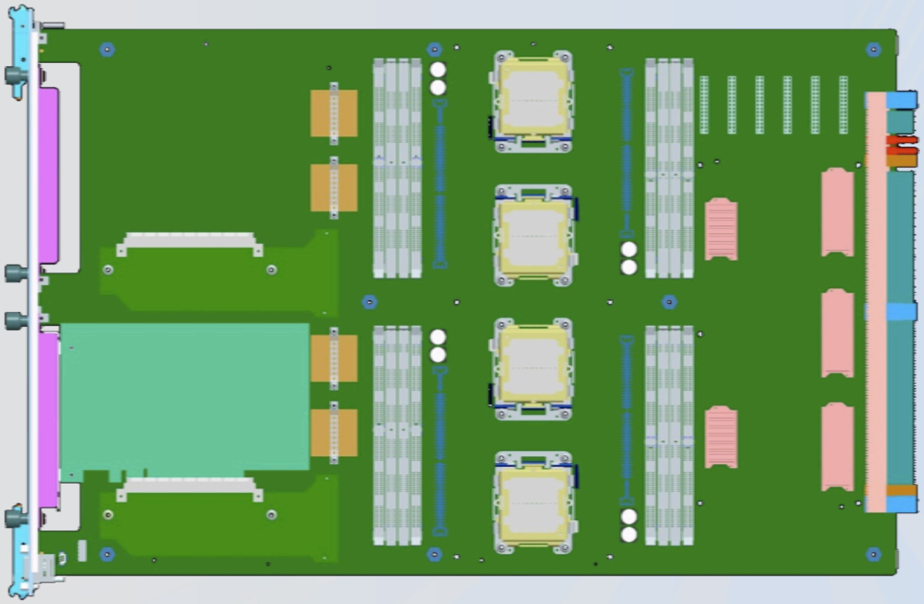
Cray I/O

will be covered Wednesday

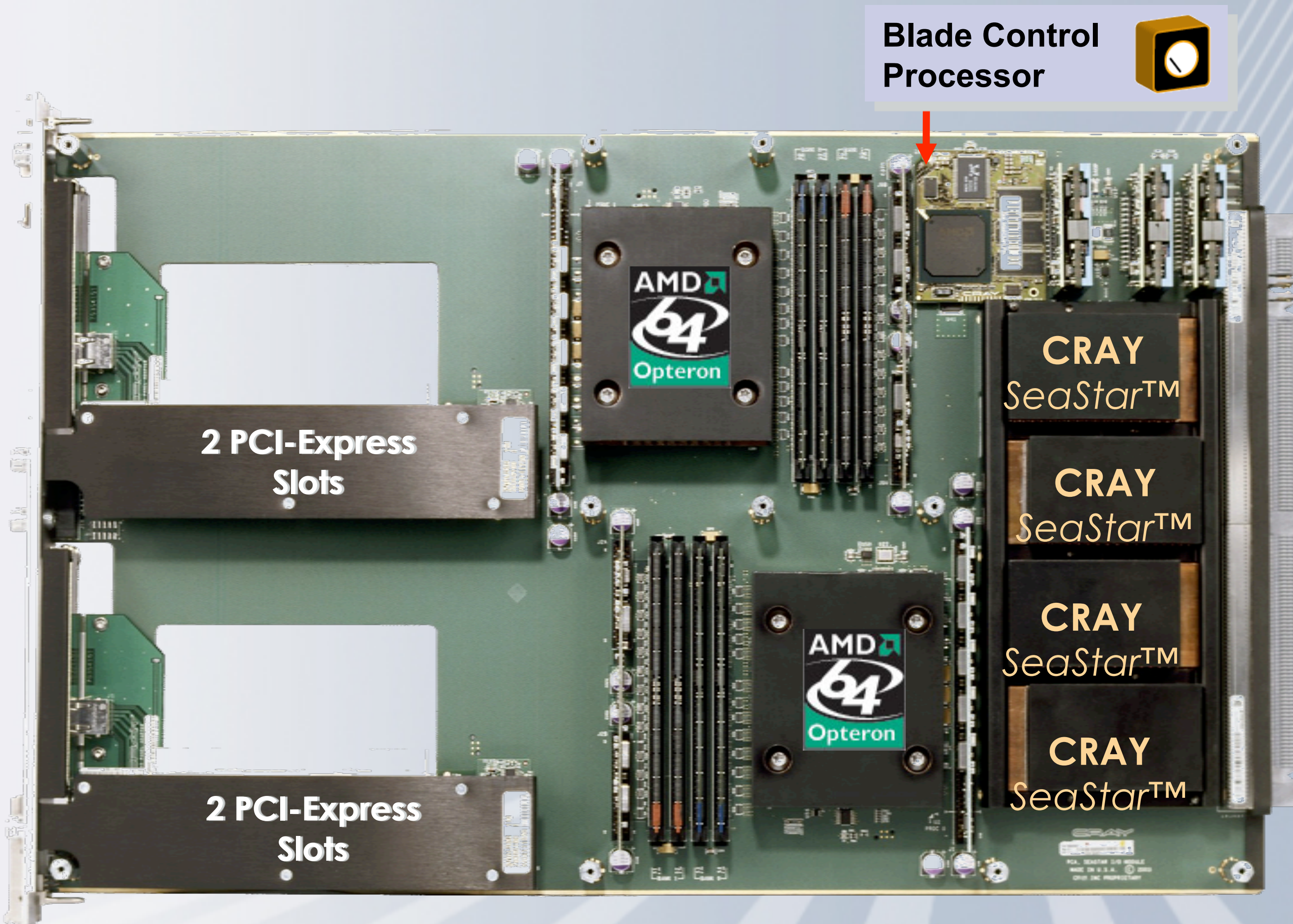
DMI Configuration – Production Weather



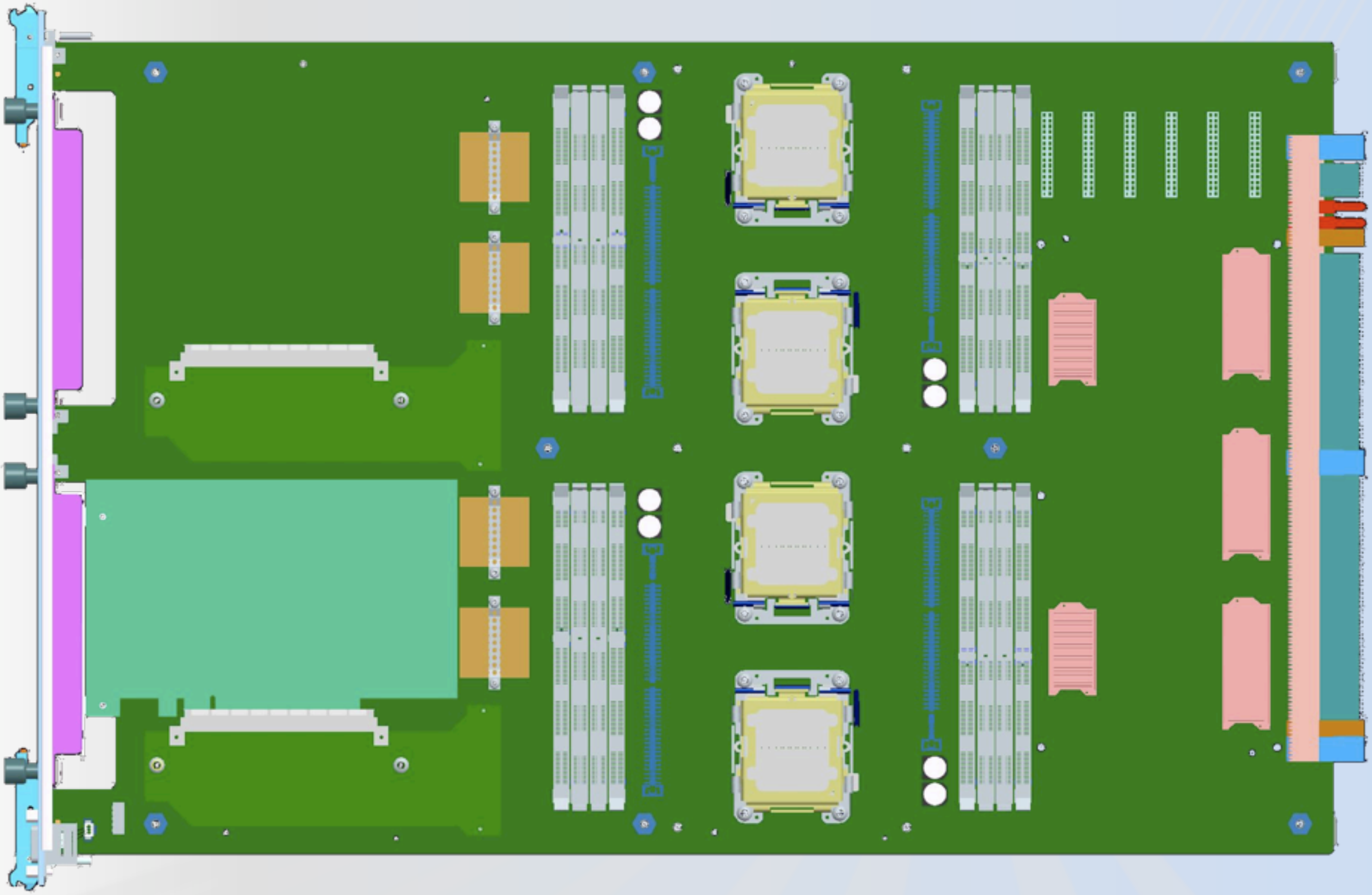
Fiorano SIO Blade



Current Service and I/O Blade

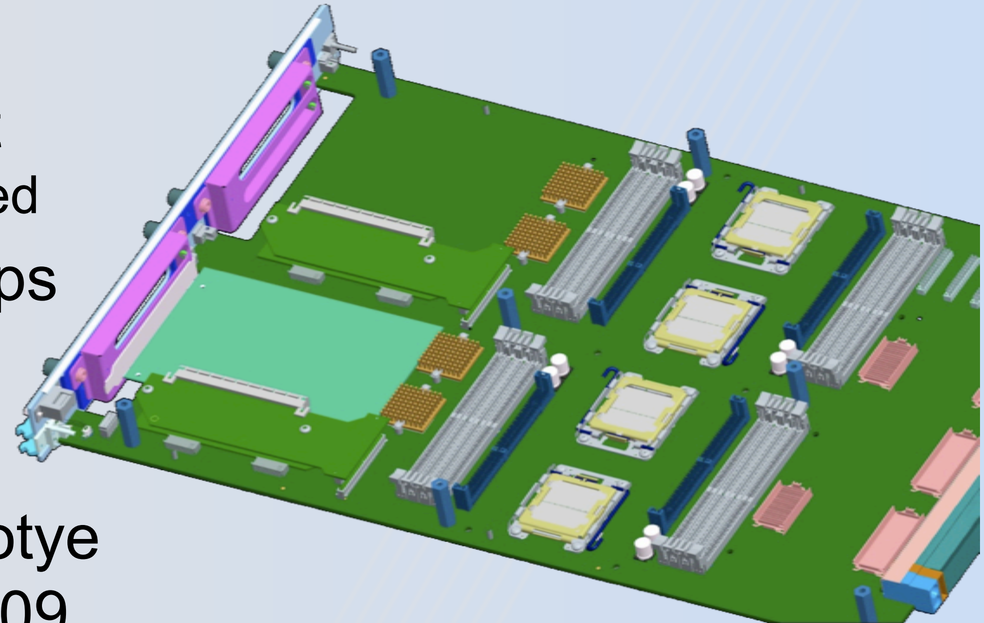


Cray “Fiorano” SIO Blade



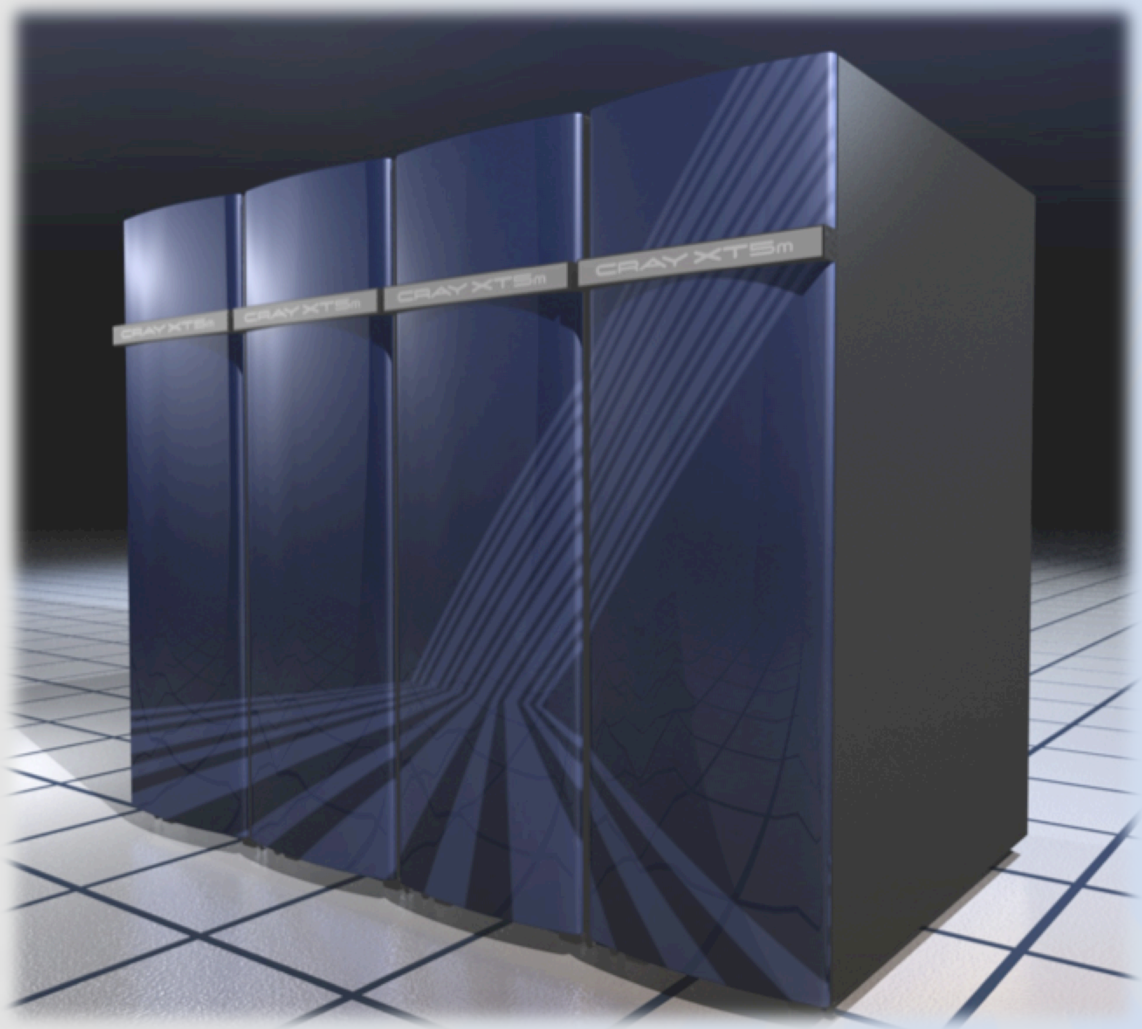
Next Generation Fiorano SIO Blade

- 4 Socket F Processors
- 4 DDR2 DIMMs per Socket
  - 4GB or 8GB DIMMs supported
- 4 AMD SR5670 Bridge Chips
- Supports both SeaStar and Gemini networks
- PC Boards in fab, first prototye power-up expected late 1Q09



Blade Feature	Current SIO Blade	Fiorano SIO Blade	Fiorano Difference
# of cores	4	16-24	6x
Max. memory size	16GB	128GB	8x
Memory Bandwidth	12.8 GB/s	51.2 GB/s	4x
Sustained I/O Bandwidth	4GB/s	16GB/s	4x

CRAY XT5m



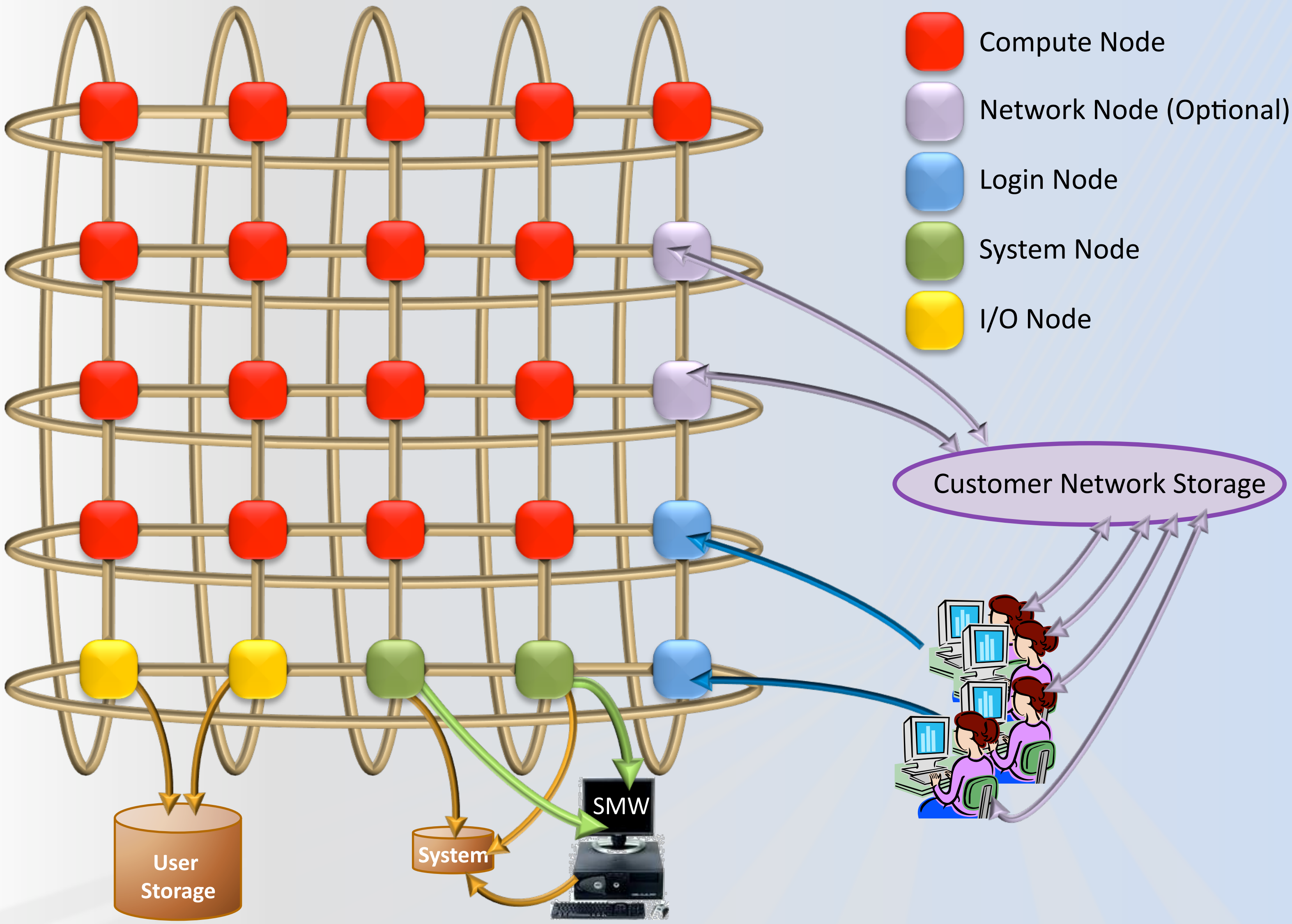
What is an XT5m

The XT5m project targets XT5 technologies at the sub $1M market

Characteristics:

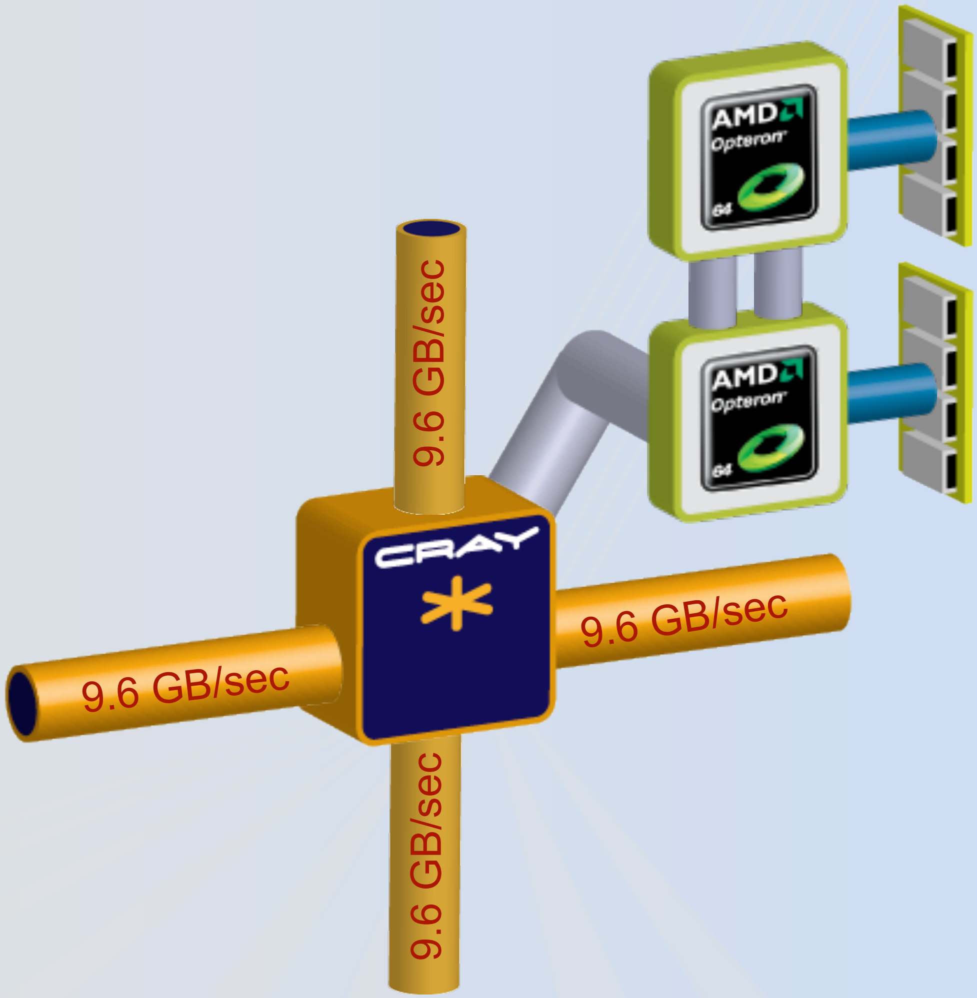
- Lower Price
- Smaller and more limited configurations
  - 1-6 Cabinets
- "Right-sized" Interconnect
  - SeaStar 1.2 Interconnect
  - 2D Torus Topology
- New "Customer Assist" Service Plan

Cray XT5m Architecture



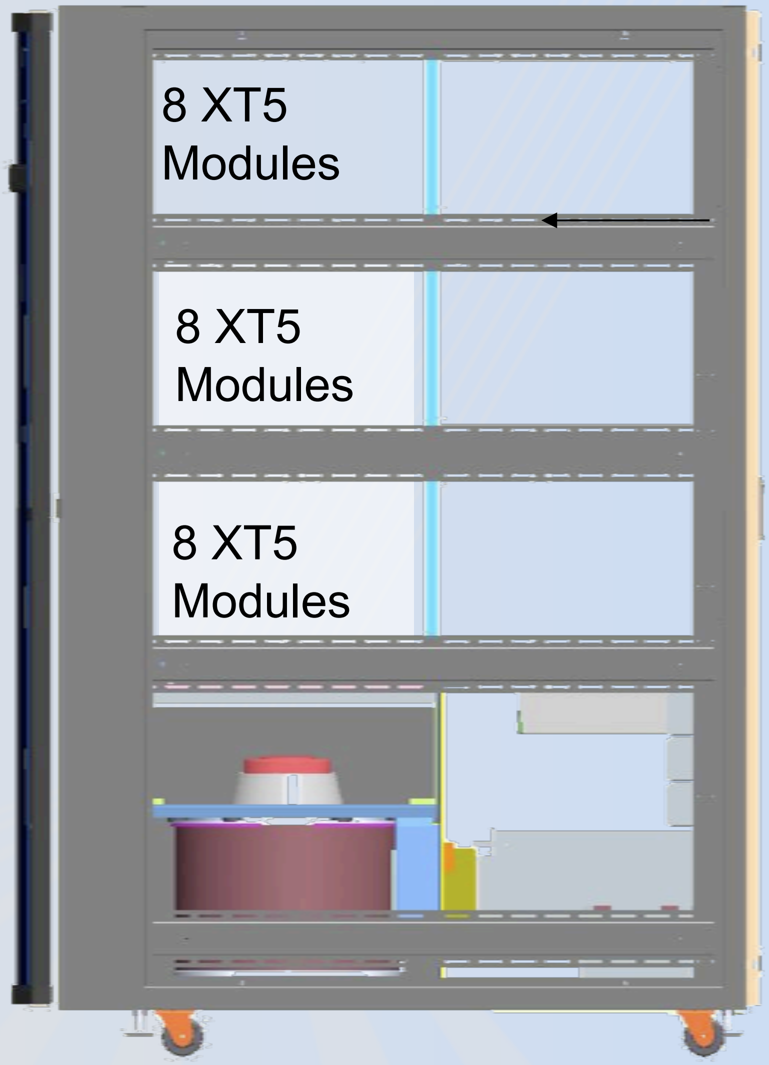
Cray XT5m Node

Cray XT5m Node Characteristics	
Number of Cores	8 or 12
Peak Performance Shanghai (2.4)	76 Gflops/sec
Peak Performance Istanbul (2.2)	105 Gflops/sec
Memory Size	8-32 GB per node
Memory Bandwidth	25.6 GB/sec



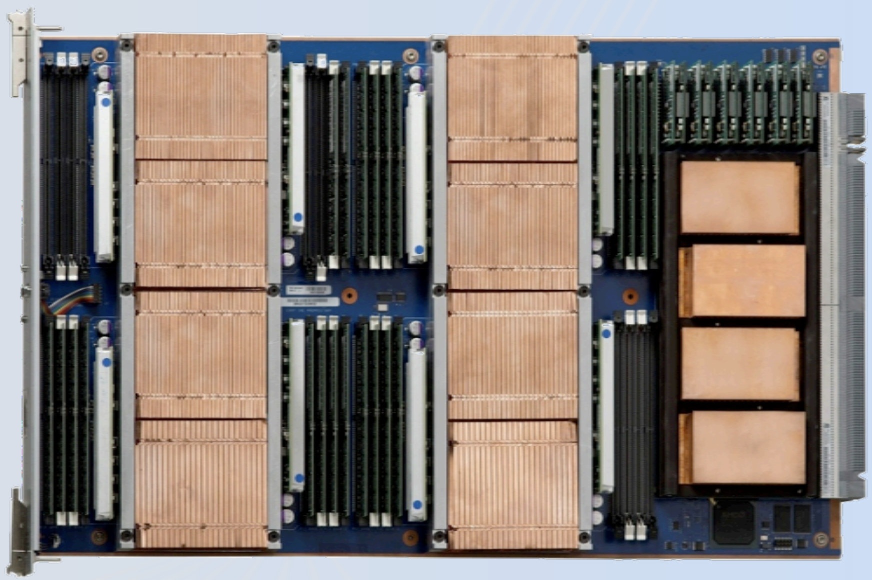
Packaging

- System is packaged in a standard XT "HE" cabinet
- 400/480 Volt PDU
- Floor Plenum required for air cooled cabinets
- 3 Chassis, 24 modules per cabinet
- ECOphlex liquid cooling is an option



Compute Blades

- Compute blades are identical to what can be installed in XT5
- Quad-core "Shanghai" processors supported
- Support for 6-core "Istanbul" in mid 2009
- Configured with SeaStar 1.2 Mezzanine cards
- 4 to 16 GB of memory per socket



XT5m Benchmark Results HPCC

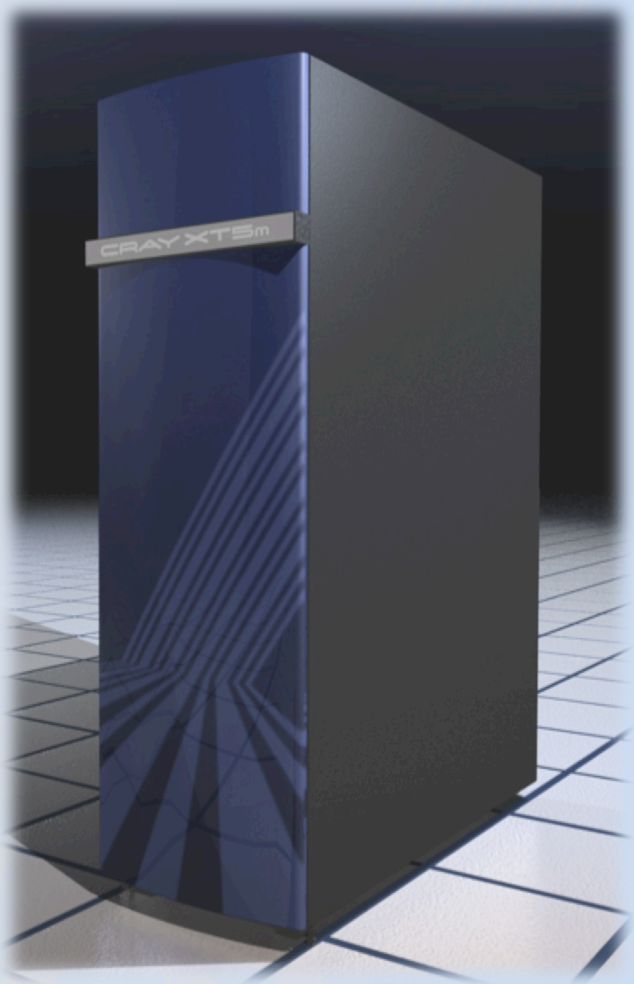
Test (672 cores)	XT5 (2.3 Barcelona)	XT5m (2.5 GHz Shanghai)
PTRANS	32.7 GB/sec	30.3 GB/sec
G Random Access	.41 GUPs	.47 GUPs
G-FFT	75.5 Gflops/sec	74 Gflops/sec
Random Ring Latency	36.3 us	36.6 us
Random Ring Bandwidth	.052 GB/sec	.048 GB/sec
Natural Ring Latency	7.6 us	7.0 us
Natural Ring Bandwidth	.51 GB/sec	.33 GB/sec

- Note that XT5m is benchmarked here with the faster Shanghai processor
- Results are very close, even for communication intensive kernels such as PTRANS and G-FFT
- Natural Ring Bandwidth benchmarks shows the difference in injection bandwidth

Single Cabinet XT5m

SPECIFICATIONS

Compute cabinets:	1 (3 chassis)
Compute Sockets:	168
Compute Cores:	672 - 1008
Peak:	**6.2 – 8.8** Tflops
Memory:	.6 – 2.6 TBytes
Topology:	12 x 8
Floor space:	2 Tiles
System power:	40 kW



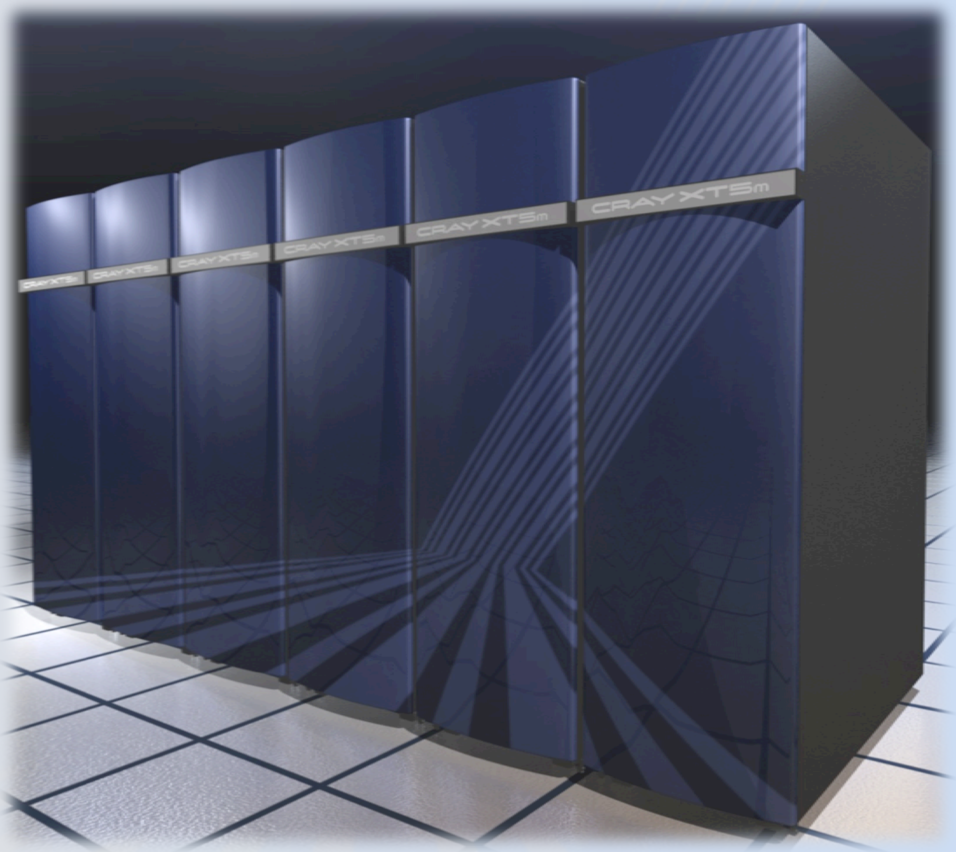
Four Cabinet XT5m

SPECIFICATIONS	
Compute cabinets:	4 (12 chassis)
Compute sockets:	736
Compute cores:	2944 - 4416
Peak:	**27 - 39** Tflops
Memory:	2.9 -1.5 TBytes
Topology:	16 x 24
Floor space:	8 Tiles
System power:	160 kW

Six Cabinet XT5m (maximum configuration)

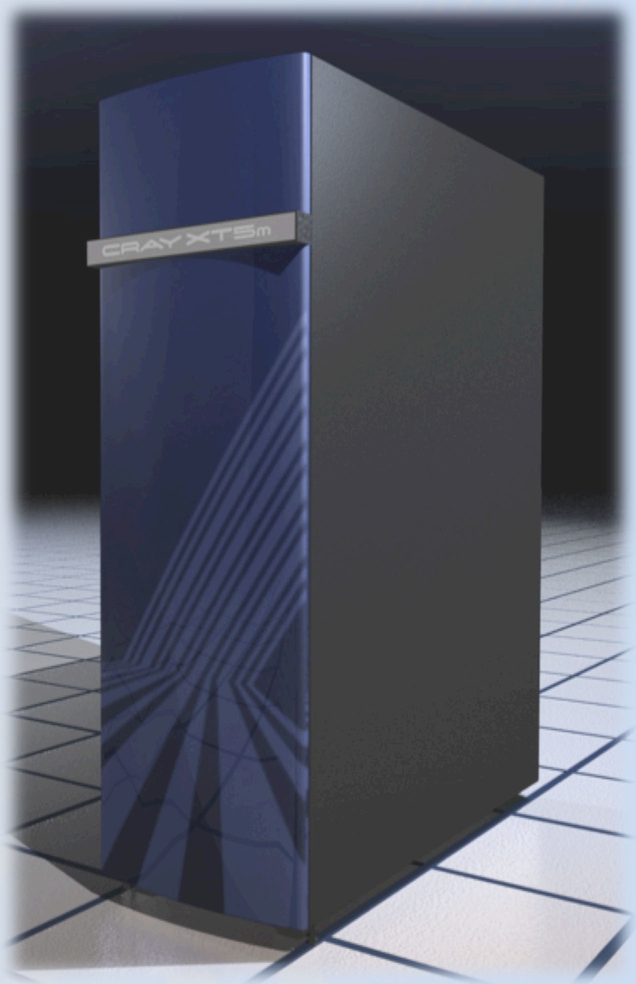
SPECIFICATIONS

Compute cabinets:	6 (18 chassis)
Compute sockets:	1120
Compute cores:	4480 - 6720
Peak:	**43 – 59** Tflops
Memory:	4.3-17.2 TBytes
Topology:	24 x 24
Floor space:	12 Tiles
System power:	240kW



XT5m - Summary

- Plan to ship in Q1 2009
- 1- 6 cabinet systems
- Upgradeable to 6 core Istanbul in mid 2009
- Upgradeable to future Baker technologies
- Cray ECOphlex liquid cooling option available
- New Customer Assist Service plan where appropriate
- Also looking at select ISV codes that make sense



The Cray High Efficiency Cabinet with ECOphlex

This is Cray's 7th method of Liquid Cooling

- In the past, liquid cooling was used primarily to increase performance
  - The game was to pack circuitry as tightly as possible
  - Or to run at higher clock cycles
  - Or to cool very high power parts
- The move to distributed programming and commodity components has reduced the need to pack things tightly

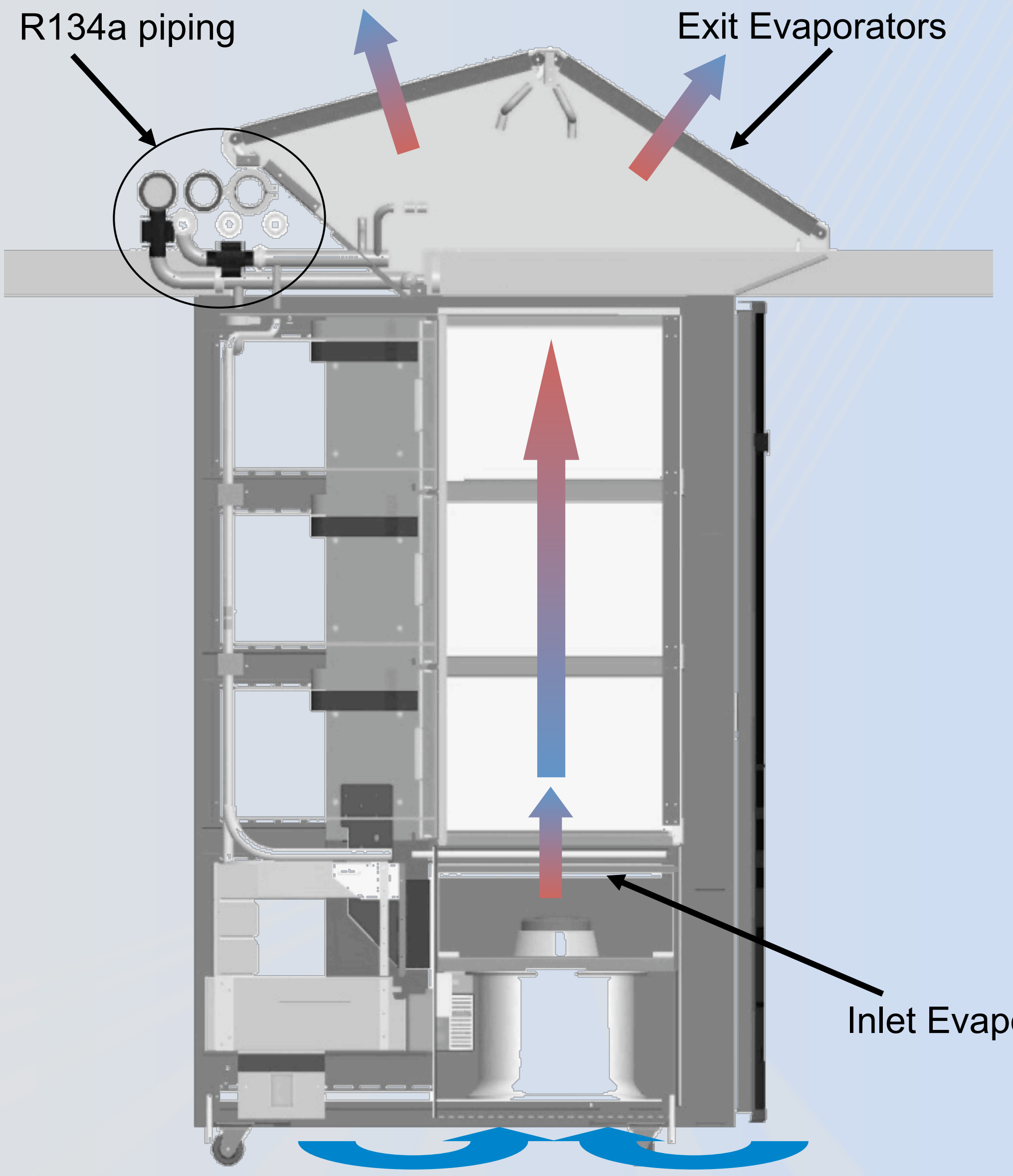
- Today, the motivations are largely based on cost of ownership
  - although the frequency "reach" of copper cables requires density

Method #7 – Cray ECOphlex cooling

ECOphlex ?? What does this mean?

- "ECO" – Stands for ECOnomical and ECOlogical
- "phlex" – Stands for PHase Liquid EXchange

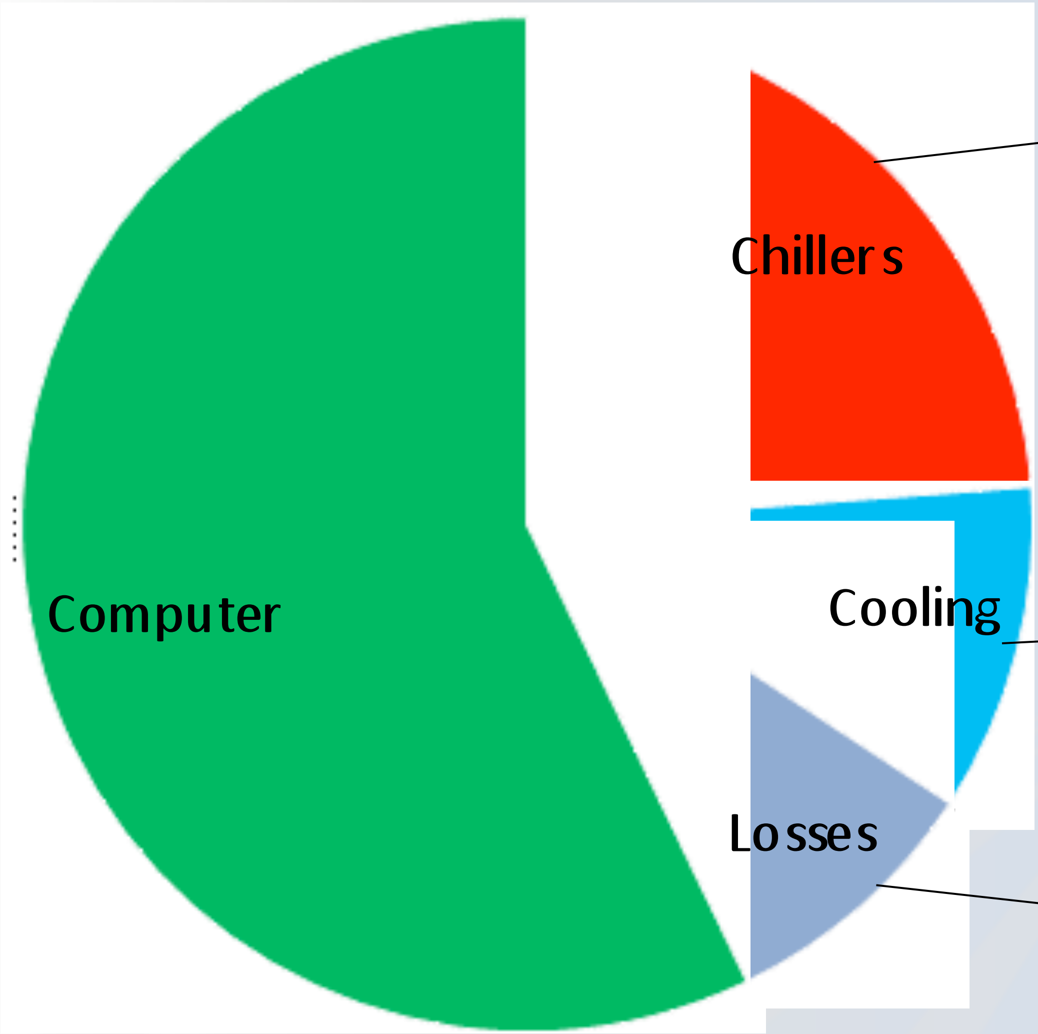
ECOphlex Technology in the Cray High Efficiency Cabinet



Advantages vs. Other Water-Cooled Solutions

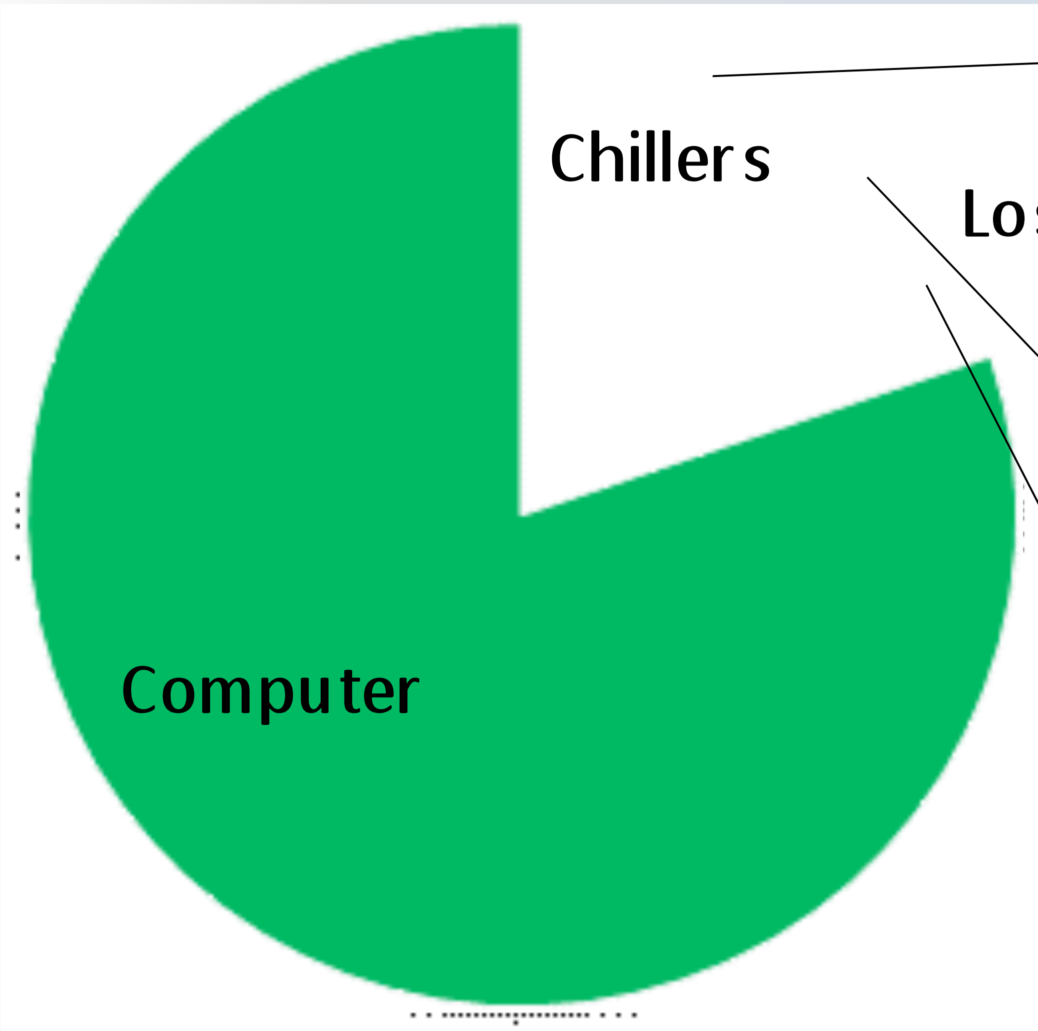
- Cost of Ownership
  - Coolant can be re-condensed using building water.
  - In many cases, cooling can be "free"
- No water near computer components
  - If leaked, will not damage components
  - No condensate drains or chance of water damage
- Lightweight
  - Small volume of coolant in the compute cabinet.
  - Floor load is similar to air-cooled cabinets
- Serviceability and Reliability
  - Blades are still air cooled and can be pulled while system in operation.
  - Large systems can continue to fully function if an HEU is down

Typical Data Center Power Efficiency



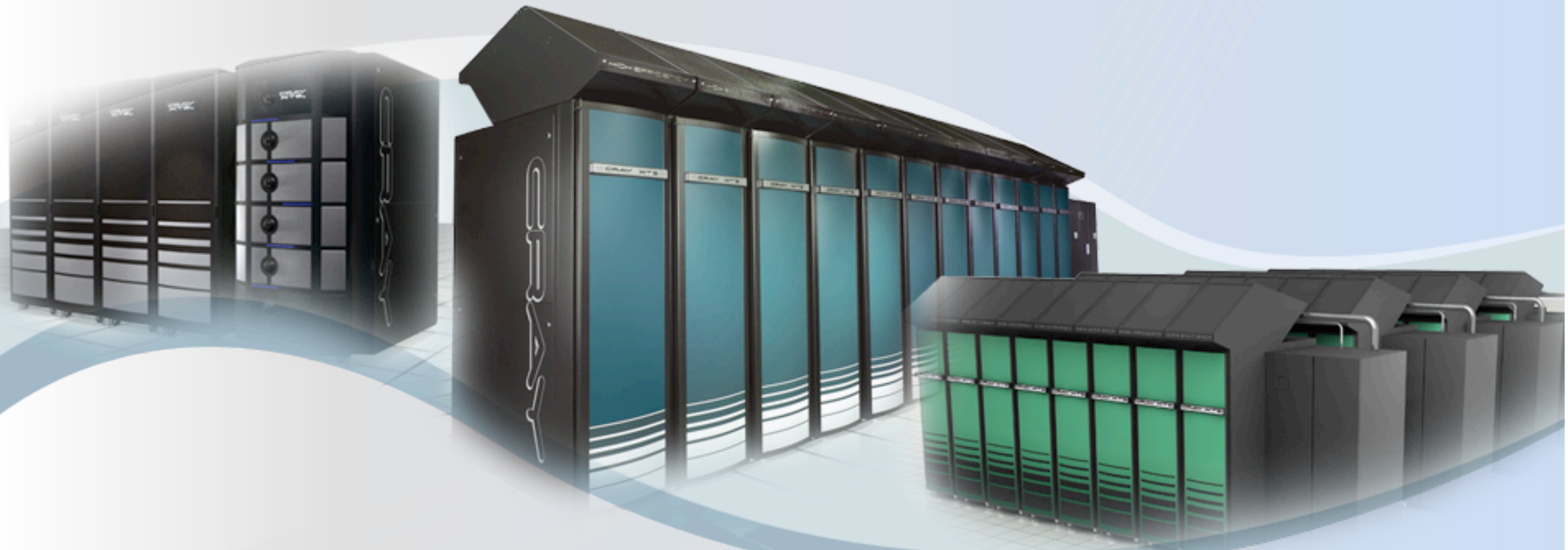
Power Conversion Losses

Cray ECOphlex Data Center Efficiency

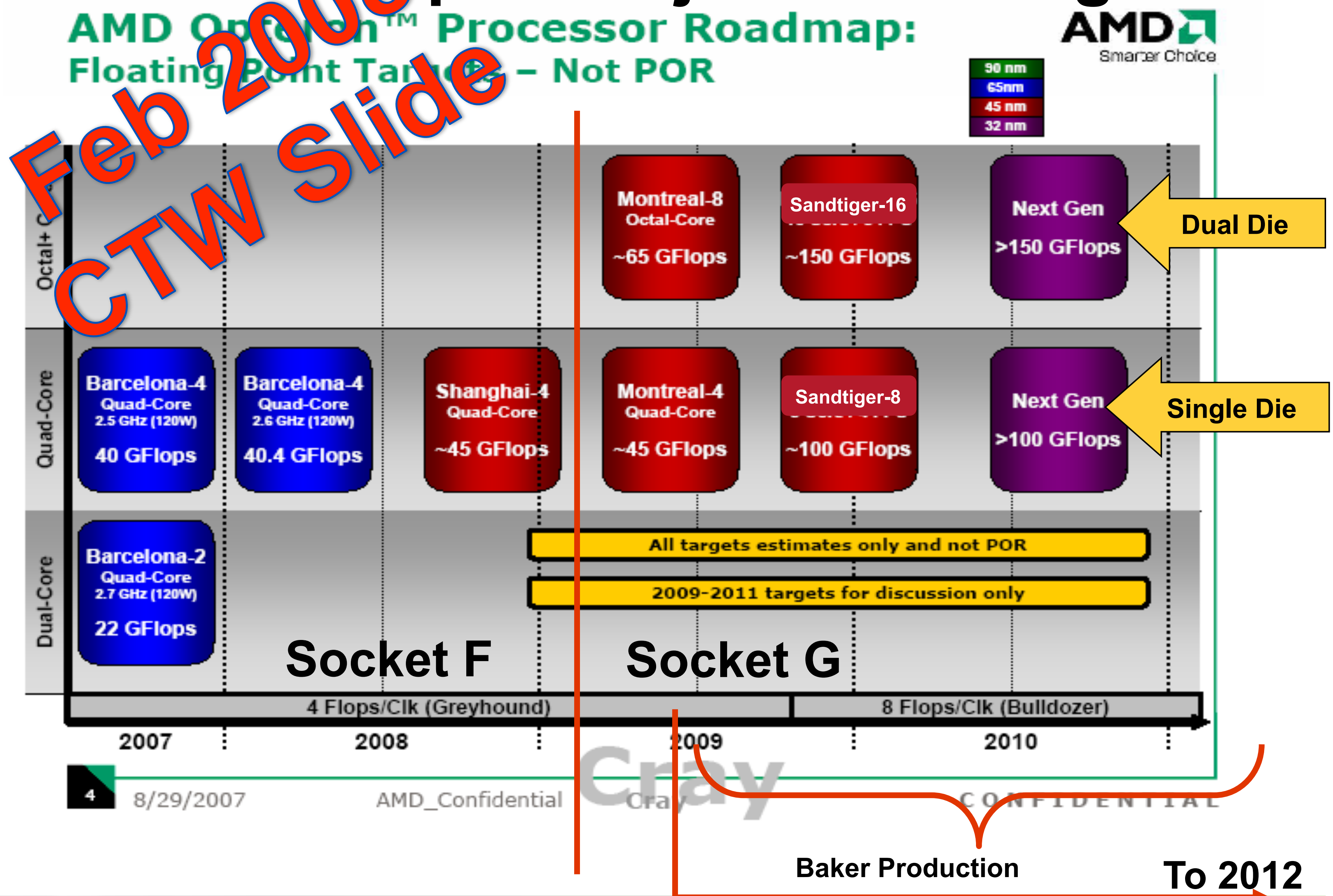


Less
Conversion
Loss (92%)

Future MPP Systems



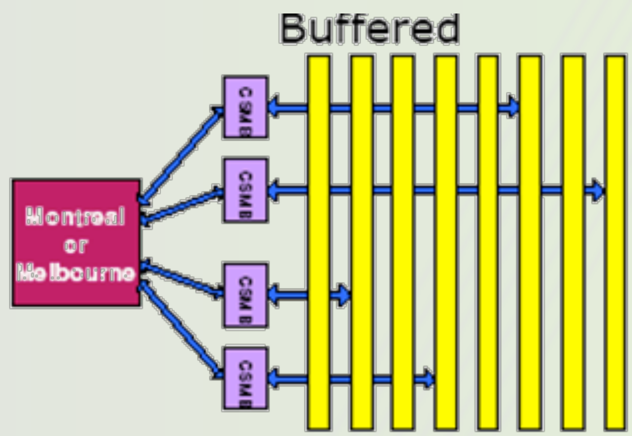
AMD Roadmap – Subject to Change



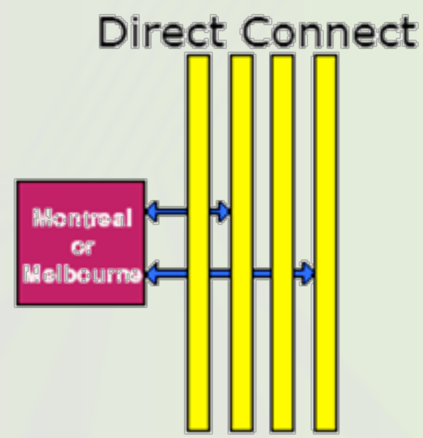
Buffered vs. Direct Connect

Feb 2008 CTW Slide

- Buffered connection requires 4 G3MX chips for each Opteron Socket
- More than one DIMM can be put behind each buffer chip
- Peak bi-directional bandwidth is 31.9 GB/sec (1333 Mhz DDR3) with a dual-die part or 21.3 GB/sec with a single-die part



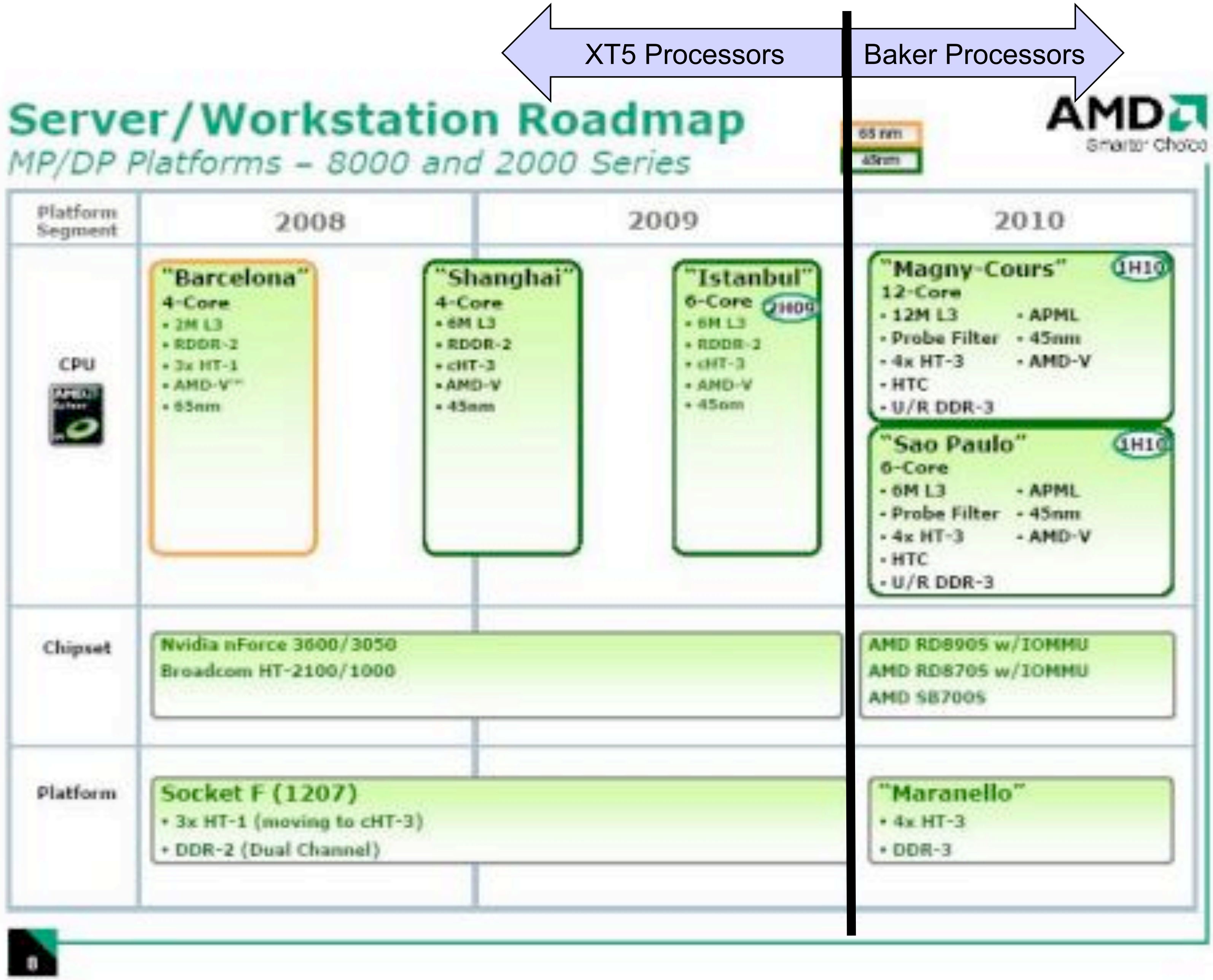
- Direct connect memory allows for two channels
- Up to two registered DIMMs per channel
- Peak bi-directional bandwidth is 21.3 GB/sec (1333 Mhz DDR3) for both single and dual-die parts



Comparison

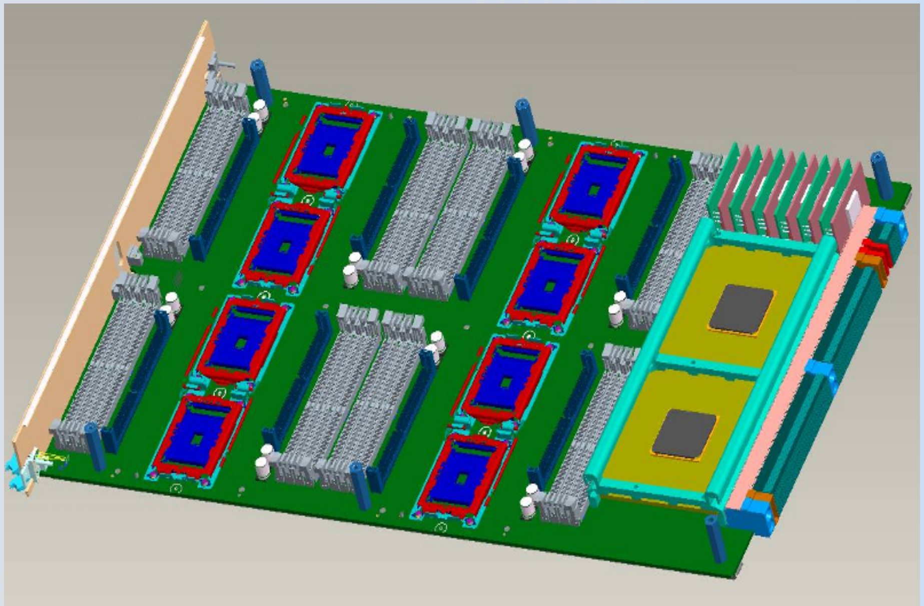
Feb 2008
CTW Slide

Attribute	Buffer Chips	Direct Connect
Nodes Per Board	4	4
Sockets Per Board	4	8
Opteron Die Per Board	4-8	8-16
Dimms Per Board	32	32
Buffer Chips per board	16	0
Peak bi-directional memory bandwidth / board	85 - 127 GB/sec	170 GB/ sec
Peak Gflops per board	400-600	800-1200

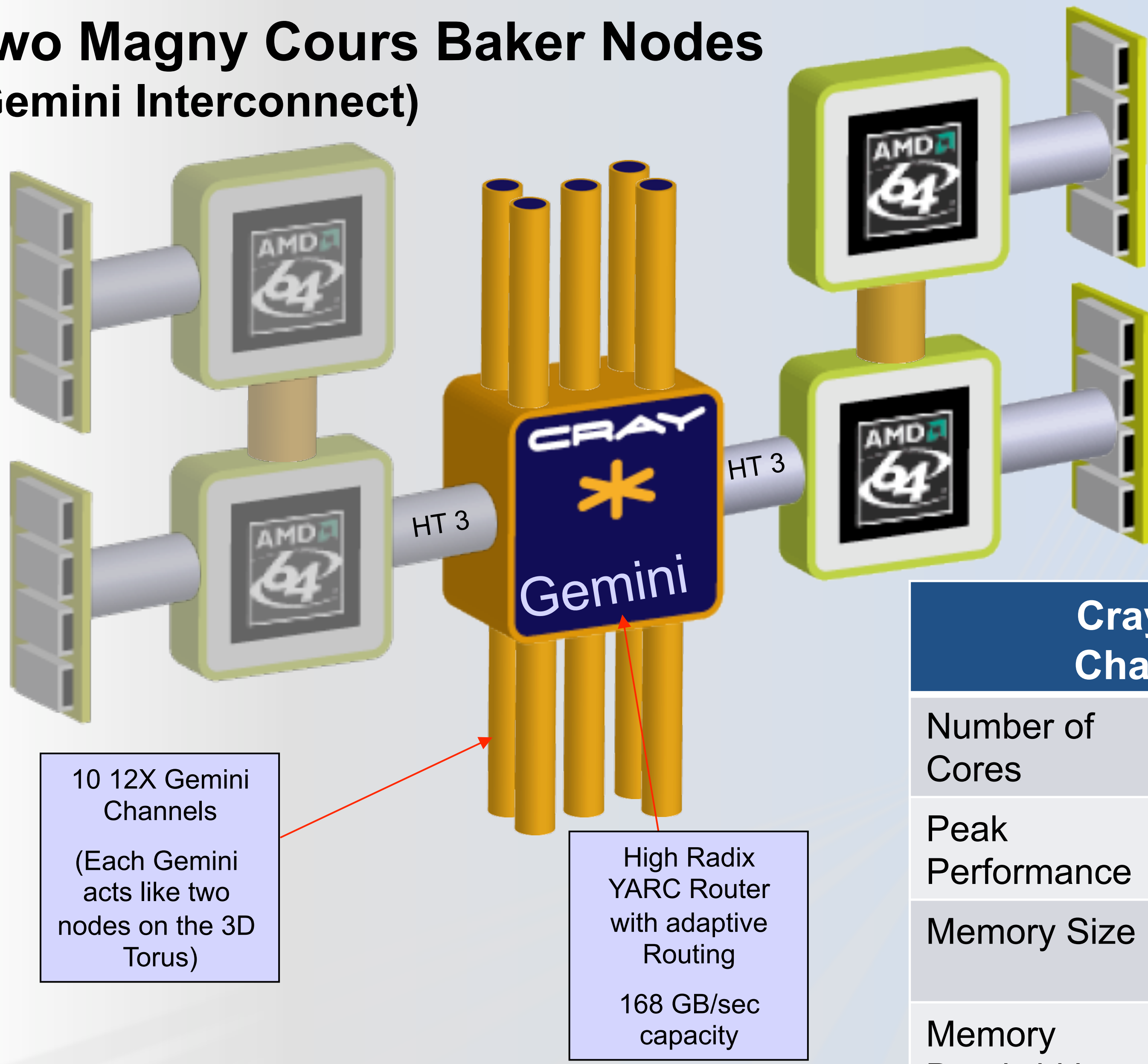

XT5 Processors
Baker Processors
Server/Workstation Roadmap
MP/DP Platforms – 8000 and 2000 Series
AMD
Platform Segment
2008
2009
2010
CPU
"Barcelona"
4-Core
"Shanghai"
4-Core
• 6M L3
• RDDR-2
• cHT-3
• AMD-V
• 45nm
"Istanbul"
6-Core
"Magny-Cours"
1H10
12-Core
• 12M L3
• Probe Filter
• 4x HT-3
• HTC
• U/R DDR-3
• APML
• 45nm
• AMD-V
"Sao Paulo"
1H10
6-Core
• 6M L3
• Probe Filter
• 4x HT-3
• HTC
• U/R DDR-3
• APML
• 45nm
• AMD-V
Chipset
Nvidia nForce 3600/3050
Broadcom HT-2100/1000
AMD RD8905 w/IOMMU
AMD RD8705 w/IOMMU
AMD SB700S
Platform
Socket F (1207)
• 3x HT-1 (moving to cHT-3)
• DDR-2 (Dual Channel)
"Maranello"
• 4x HT-3
• DDR-3

Baker Compute Blade – 2H 2010

- 8 AMD Socket G34 Processors
- Initially targeting 12-core Magny-Cours Processor
- 4 DDR3 DIMM slots per socket
- *4 Channels full DDR3 Bandwidth per socket*
- Compatible with SeaStar or Gemini Mezzanine card

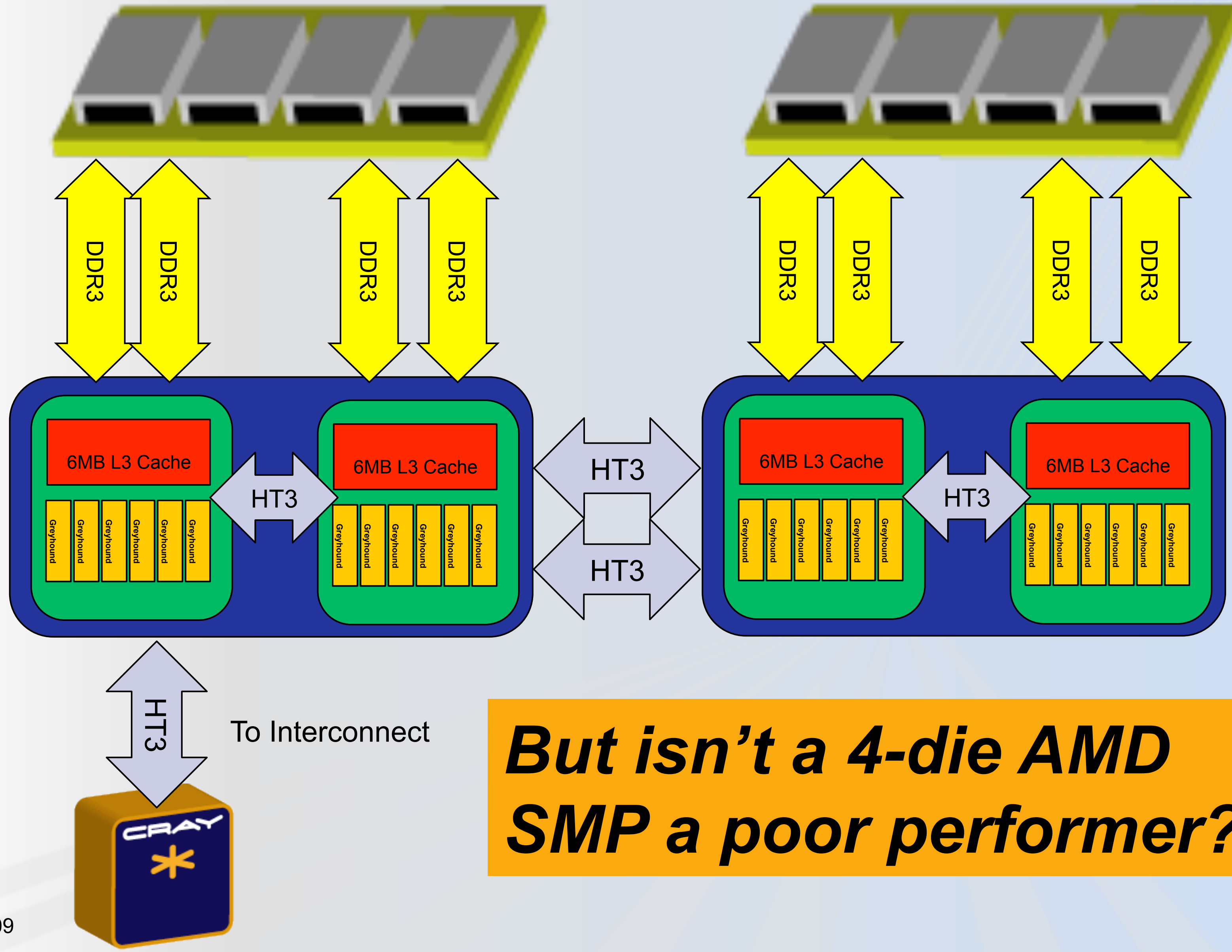


Two Magny Cours Baker Nodes
(Gemini Interconnect)



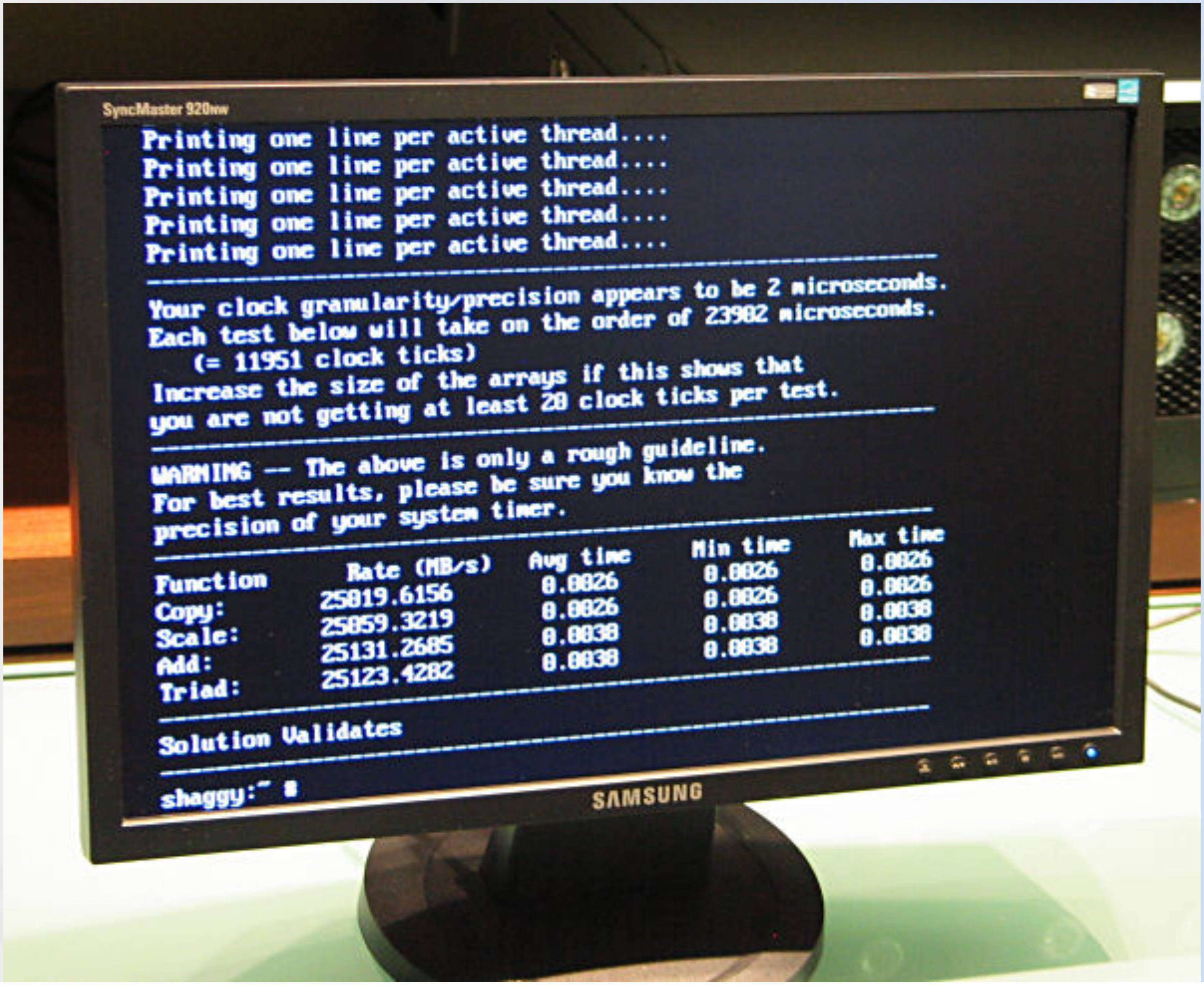
Cray XT5 Node Characteristics	
Number of Cores	24
Peak Performance	182 Gflops/s
Memory Size	32 or 64 GB per node
Memory Bandwidth	85 GB/sec

Magny Cours Baker Node (Mid 2010)



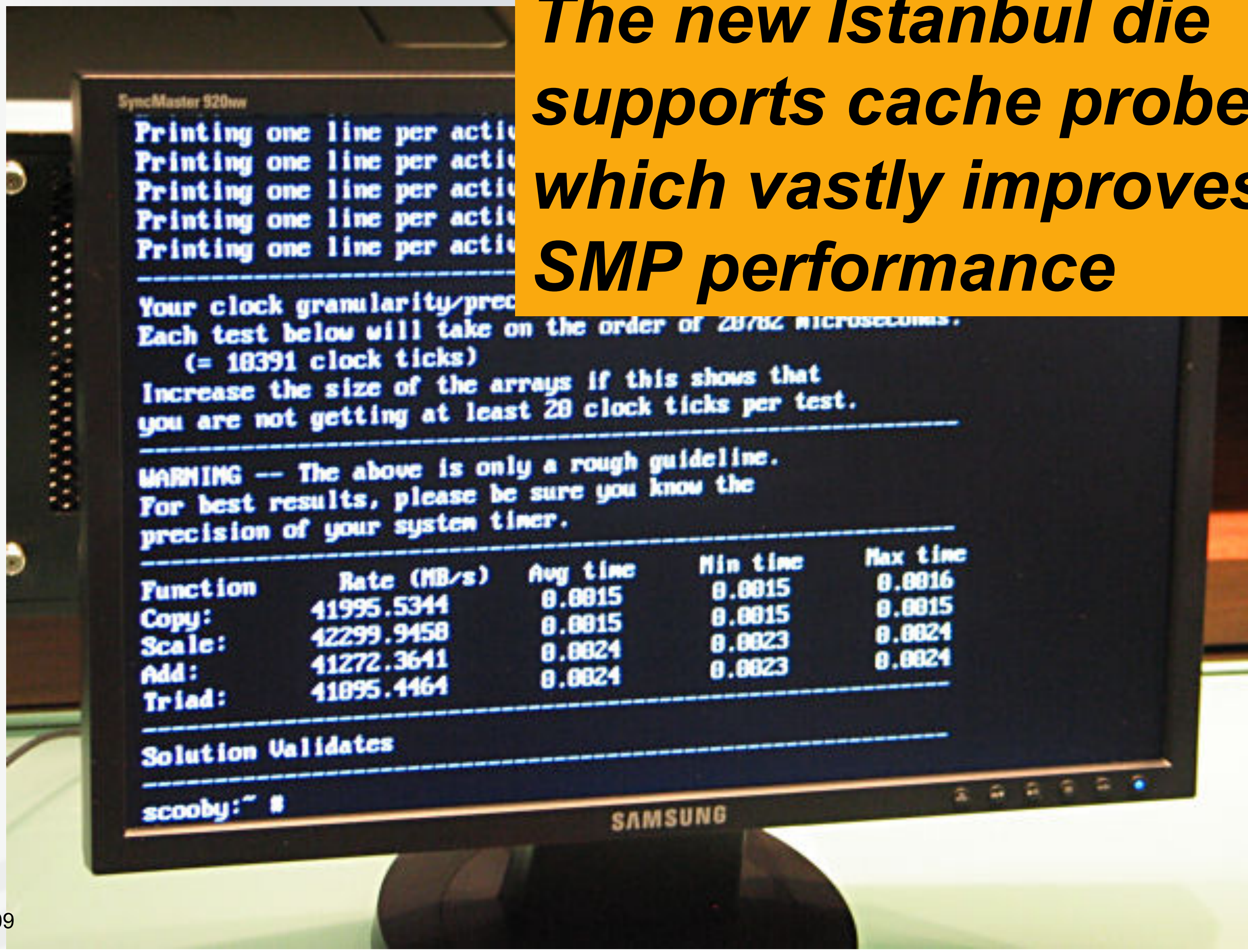
But isn't a 4-die AMD SMP a poor performer?

4-socket Barcelona Streams Test

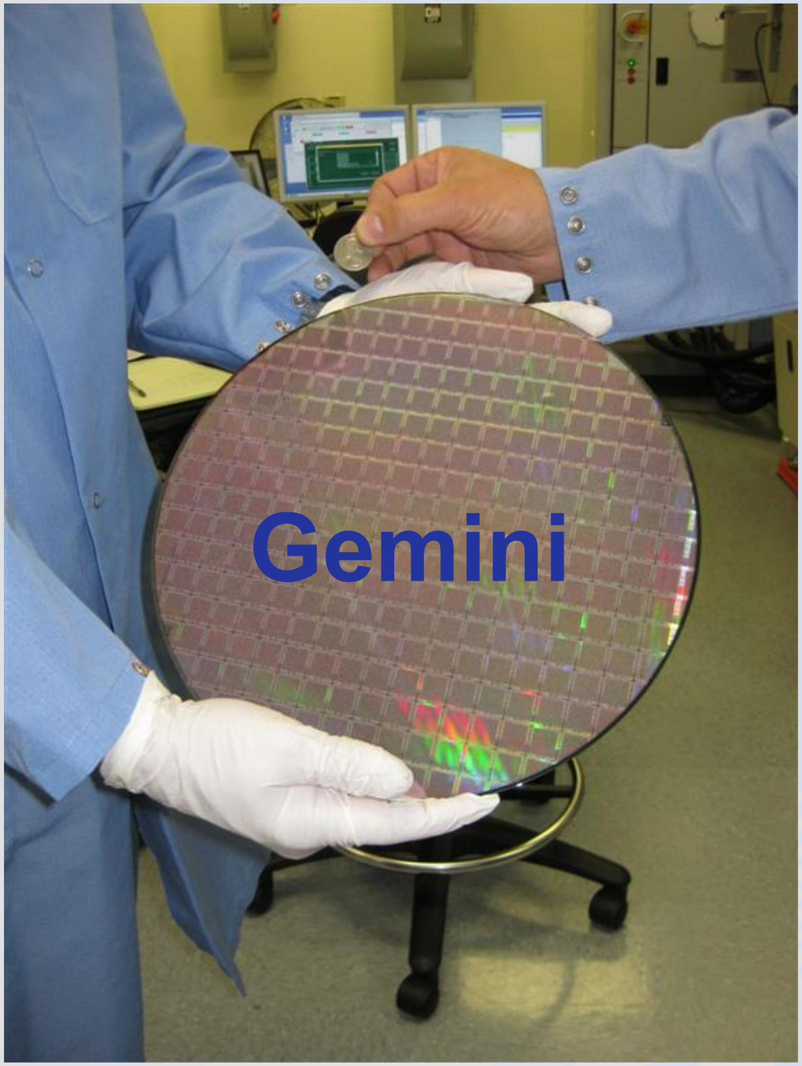


Exact same board, with 4 Istanbul

The new Istanbul die supports cache probes which vastly improves SMP performance

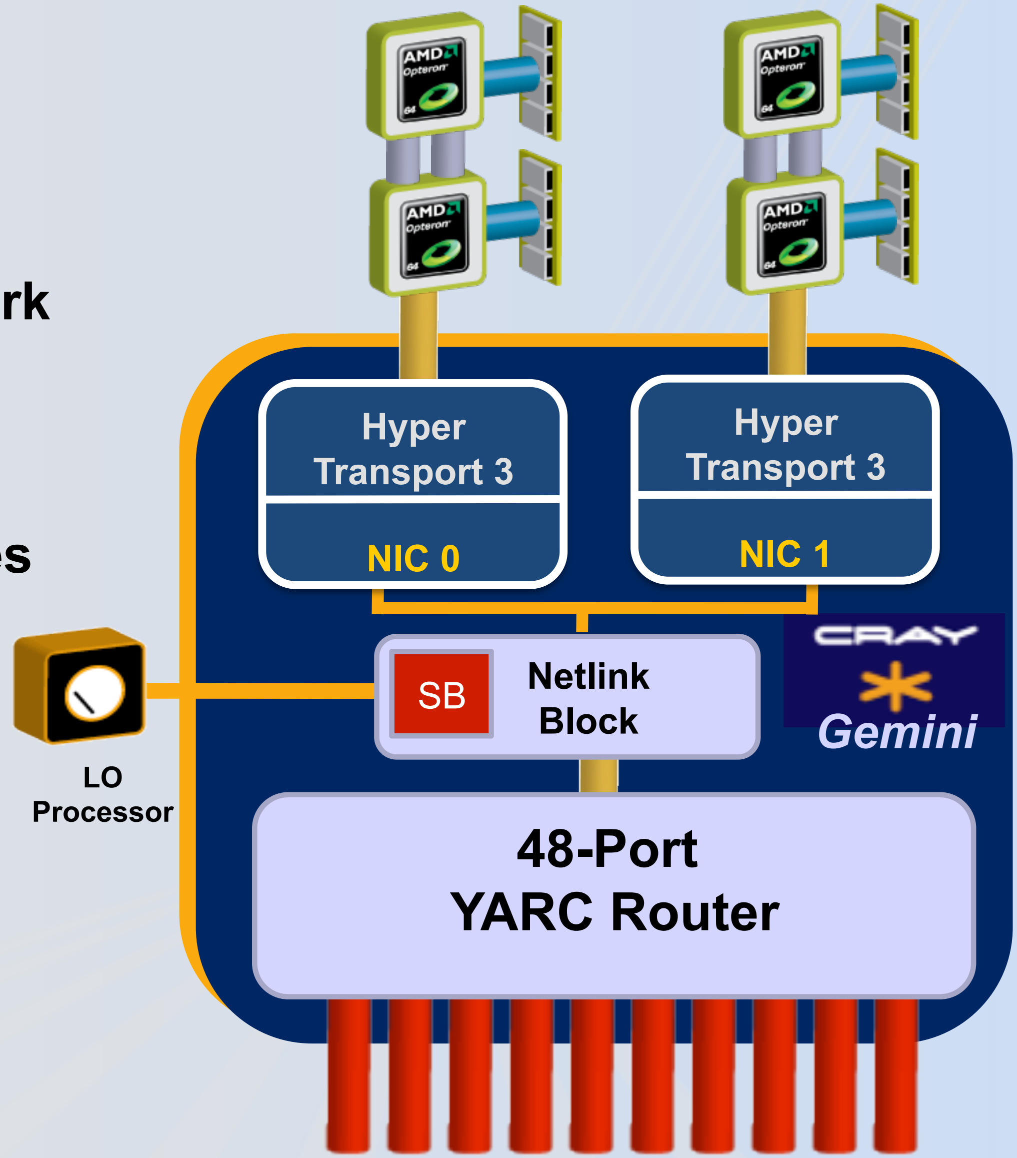


Gemini Interconnect

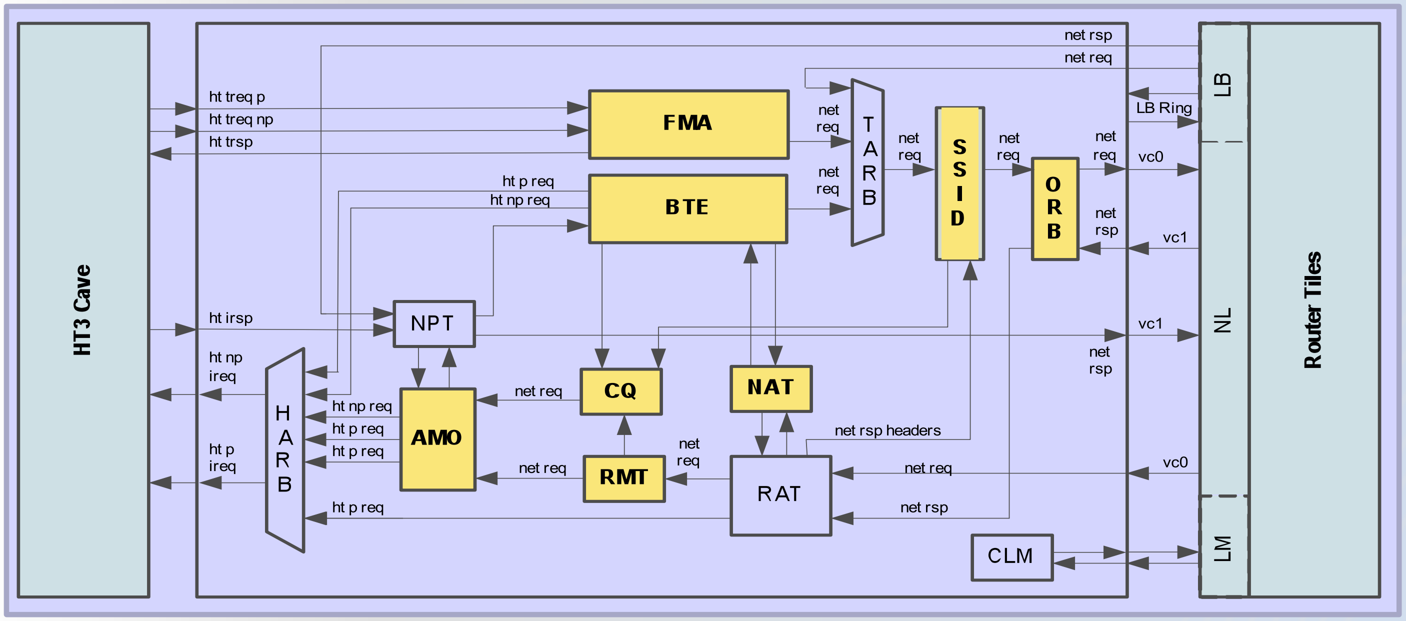


Cray Gemini ASIC

- Supports 2 Nodes per ASIC
- 168 GB/sec routing capacity
- Scales to over 100,000 network endpoints
- Link Level Reliability and Adaptive Routing
- Advanced Resiliency Features
- Provides global address space
- Advanced NIC designed to efficiently support
  - MPI
  - One-sided MPI
  - Shmem
  - UPC, Coarray FORTRAN, Titanium, Global Arrays



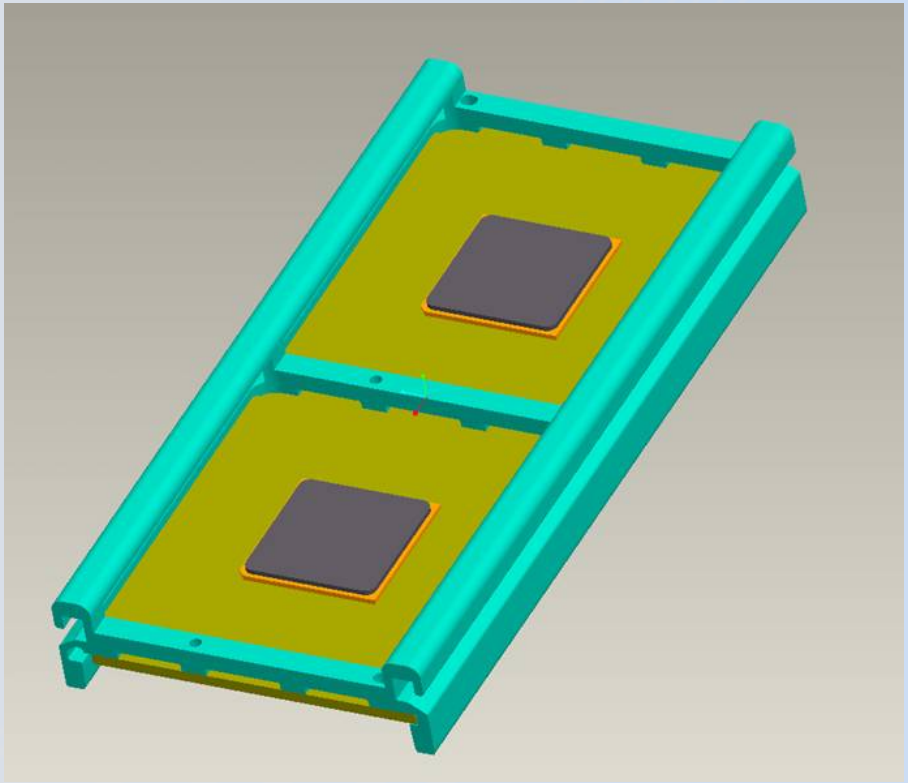
Gemini NIC block diagram



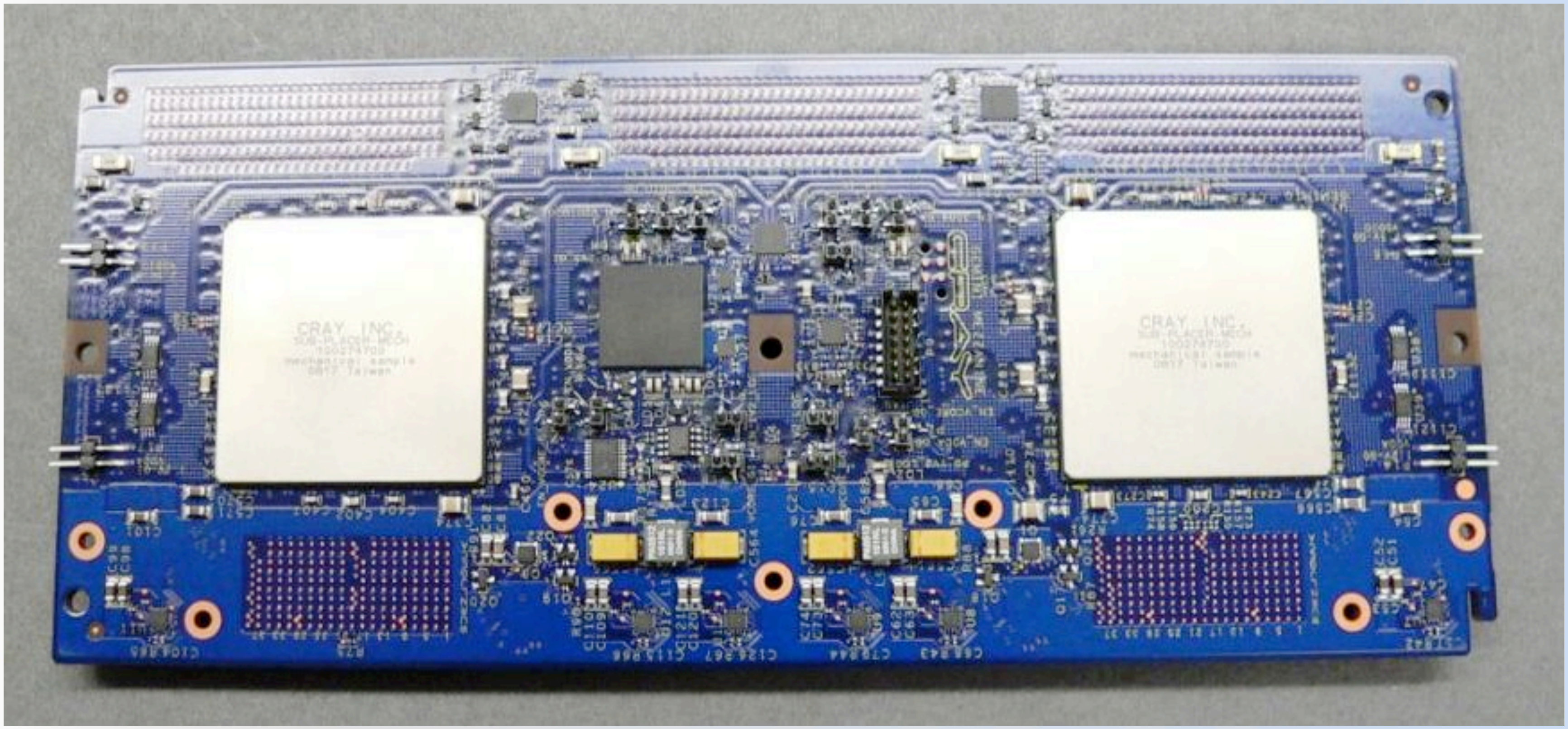
- **FMA (Fast Memory Access)**
  - Mechanism for most MPI transfers
  - Supports tens of millions of MPI requests per second
- **BTE (Block Transfer Engine)**
  - Supports *asynchronous* block transfers between local and remote memory, in either direction
  - For use for large MPI transfers that happen in the background

Gemini Mezzanine Card

- Two Gemini ASICs are packaged on a pin-compatible mezzanine card
- Topology is a 3-D torus
- Each lane of the torus is composed of 4 Gemini router "tiles"
- Systems with SeaStar interconnects can be upgraded by swapping this card
- 100% of the 48 router tiles on each Gemini chip are used



Gemini Mezzanine Card



Gemini Reliability Features

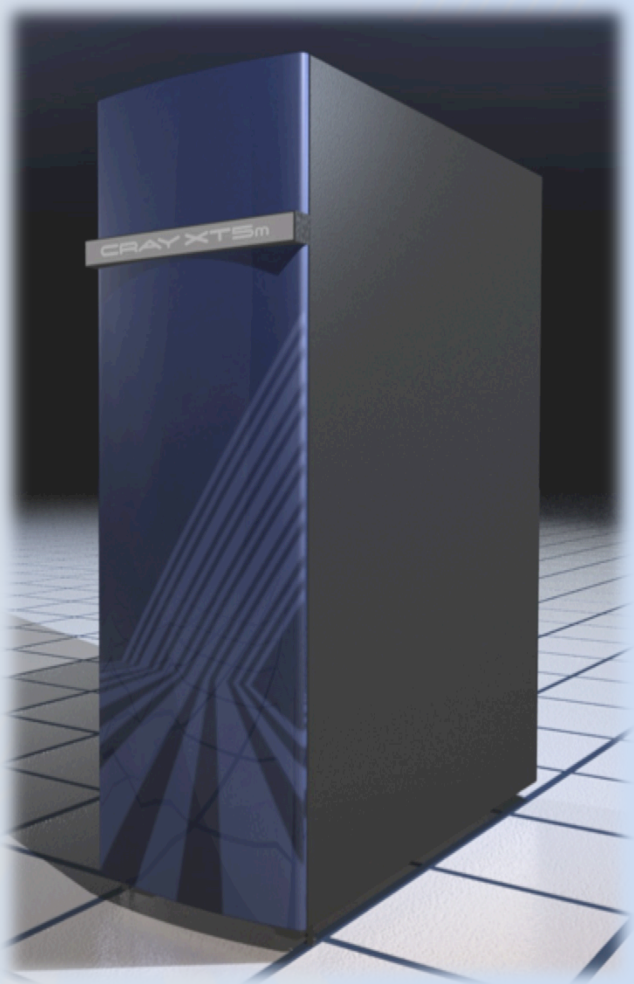
- Will support warm-swap of blades
- Can map around bad links without rebooting
- Adaptive Routing – multiple paths to the same destination
- Packet level CRC carried from start to finish
- Network channels can automatically degrade
- Large blocks of memory protected by ECC
- Can better handle failures on the HT-link, discards packets instead of putting backpressure into the network
- Improved error reporting and handling
- Performance counters allowing tracking of app specific packets
- The “send/receive” channel protocol supports end-to-end reliable communication. (used by MPICH2 and OpenMPI)
- The RDMA protocol supports low overhead verification of success or failure. The low overhead error reporting allows the programming model to replay failed transactions

Gemini – Status

- Cray approved the netlist release 8/22/08
- First Wafers out of fab on 10/25/08
- Software infrastructure in place
- First Gemini mezzanine assemblies powered up 11/17/08
- First bugs in parts found and characterized, fibbed parts returned
- First MPI message traffic on 2/10/09
  - Un-optimized, zero-byte latency between two nodes was less than 2 microseconds

Single Cabinet Baker-m

SPECIFICATIONS	
Compute cabinets:	1 (3 chassis)
Compute Sockets:	168
Compute Cores:	1512
Peak:	**15.3** Tflops
Memory:	2.6 – 5.2 TBytes
Topology:	12 x 8
Floor space:	2 Tiles
System power:	~55 kW



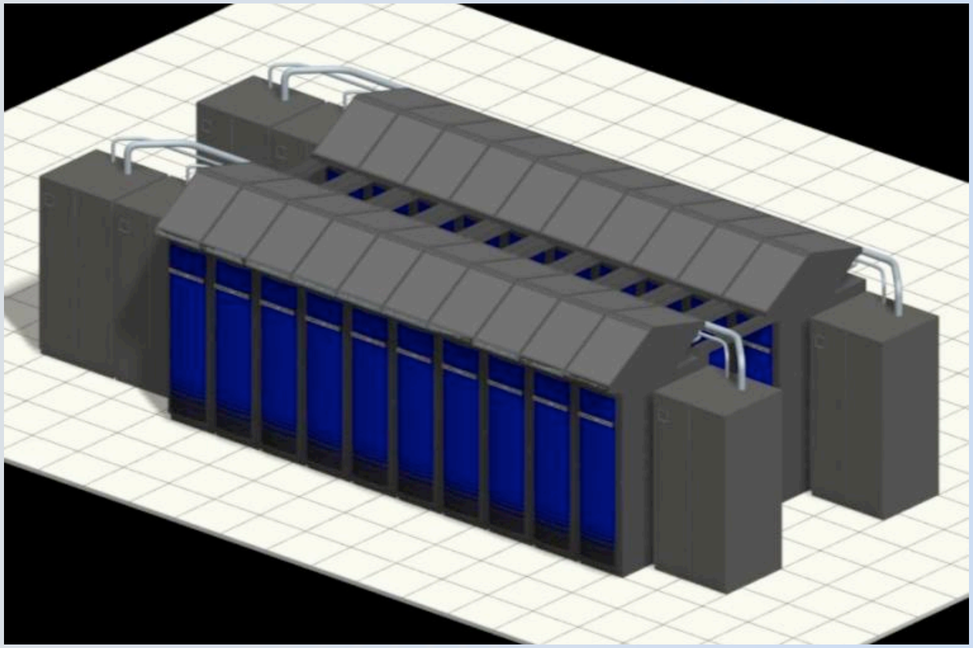
Four Cabinet Baker m

<u>SPECIFICATIONS</u>

Compute cabinets:	4 (12 chassis)
Compute sockets:	736
Compute cores:	8832
Peak:	**67.1** Tflops
Memory: TBytes	11.5 – 23
Topology:	16 x 24
Floor space:	8 Tiles
System power:	~220 kW

Example: 20 Cabinet Baker

SPECIFICATIONS	
Compute cabinets:	20 (60chassis)
Sockets:	3648
Cores:	31,776
Peak:	333 Tflops
Memory:	28 - 57 TBytes
XDP cabinets:	6 (will be 4)
Floor space:	50 Sq Meters

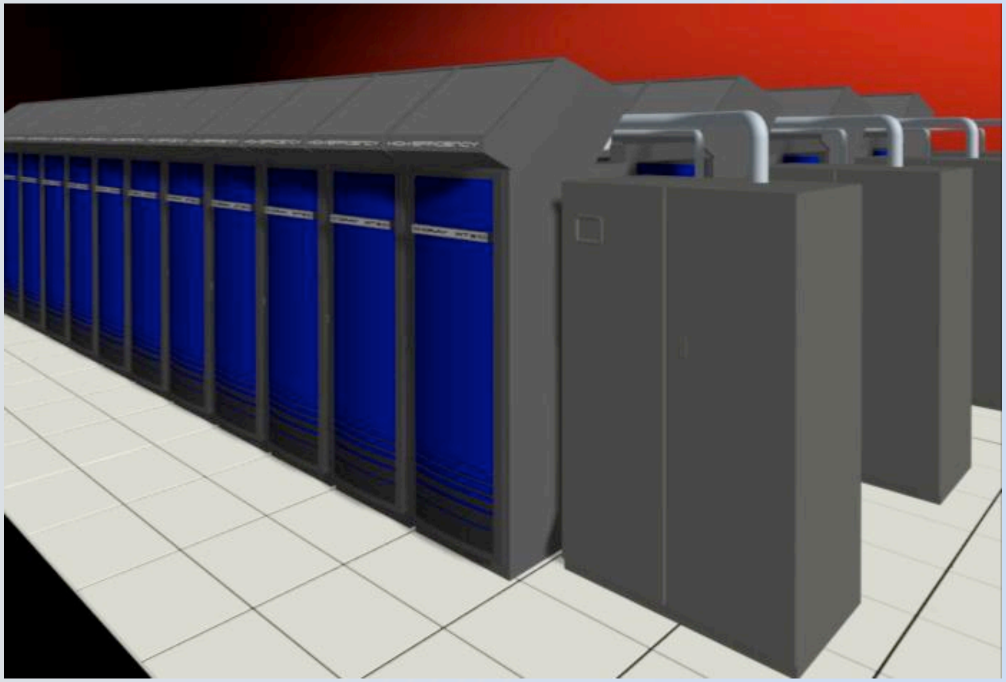


Example: 80 Cabinet Baker

SPECIFICATIONS

Compute cabinets:	80 (240 chassis)
Sockets:	15,024
Cores:	180,288
Peak:	1.36 Pflops
Memory:	117-234 TBytes
XDP cabinets:	20
Floor space:	195 Sq Meters

Floor space does not include IO & storage units



End